DOCUMENT RESUME

ED 299 130	SE 049 653

AUTHOR	Bagwell, Charles Bruce
TITLE	Theory and Application of DNA Histogram Analysis.
PUB DATE	Aug 79
NOTE	340p.; Ph.D. Dissertation, University of Miami. Contains some light and broken type.
PUB TYPE	Reports - Research/Technical (143) -- Dissertations/Theses - Doctoral Dissertations (041)
EDRS PRICE	MF01/PC14 Plus Postage.
DESCRIPTORS	Biological Sciences; *College Science; *DNA; Doctoral Dissertations; Higher Education; *Least Squares Statistics; Science Education; *Statistical Analysis; *Statistical Distributions
IDENTIFIERS	*Histograms

ABSTRACT

The underlying principles and assumptions associated with DNA histograms are discussed along with the characteristics of fluorescent probes. Information theory was described and used to calculate the information content of a DNA histogram. Two major types of DNA histogram analyses are proposed: parametric and nonparametric analysis. Three levels of parametric analysis are presented: crude stripping, pattern recognition adjustment, and multirectangle nonlinear least squares analysis. The multirectangle method's assumptions were substantiated with theoretical justifications and biological experiments. Methods of optimal smoothing, capping, translocation, and doublet correction are also discussed. The statistical comparison of groups of histograms, nonparametric analysis, is introduced and applied to the clinical problem of parathyroid disease. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

SE

ED299130

THE UNIVERSITY OF MIAMI

THEORY AND APPLICATION OF DNA HISTOGRAM ANALYSIS

by

Charles Bruce Bagwell

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

August, 1979

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Charles
Bagwell
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SE 049 653

BEST COPY AVAILABLE

THE UNIVERSITY OF MIAMI

THEORY AND APPLICATION OF DNA HISTOGRAM ANALYSIS

by

Charles Bruce Bagwell

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

August, 1979

THE UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

THEORY AND APPLICATION OF DNA HISTOGRAM ANALYSIS

Charles Bruce Bagwell

Approved:

Jo A. Jensen

Joerg A. Jensen
Professor of Microbiology
Chairman of Dissertation Committee

Laphalle Full

Laphalle Fuller
Assistant Professor of Microbiology

George L Irvin III

George L. Irvin, III
Professor of Surgery

J. H. Jett

James Jett
Staff Member, Biophysics Group.
Los Alamos Scientific Laboratory, N.M.

Clarence G. Stuckwisch
Dean of the Graduate SChool

Diana M Lopez

Diana M. Lopez
Associate Professor of Microbiology

Jerry L. Hudson
Research biologist, National
Center of Toxicological Research

Copyright, Charles Bruce Bagwell, 1979

BAGWELL, CHARLES BRUCE (M.D. ,Ph.D., Microbiology)
THEORY AND APPLICATION OF DNA HISTOGRAM ANALYSIS (August, 1979)
DNA HISTOGRAM ANALYSIS
Abstract of a doctoral dissertation at the University of Miami.
Dissertation supervised by Professor J. A. Jensen.

The underlying principles and assumptions associated with DNA histograms are discussed along with the characteristics of fluorescent probes. Information theory was described and used to calculate the information content of a DNA histogram. Two major types of DNA histogram analyses are proposed: parametric and nonparametric analysis. Three levels of parametric analysis are presented: crude stripping, pattern recognition adjustment, and multirectangle nonlinear least squares analysis. The multirectangle method's assumptions were substantiated with theoretical justifications and biological experiments. Methods of optimal smoothing, capping, translocation, and doublet correction are also discussed. The statistical comparison of groups of histograms, nonparametric analysis, is introduced and applied to the clinical problem of parathyroid disease.

ACKNOWLEDGEMENTS

There are a large number of people where without their help this work could never have been accomplished. I would like to thank Dr. Philip Marcus for initially guiding me into research. I would like to acknowledge Dr. Jerry Hudson for showing me that research could be fun as well as rewarding. I also wish to thank Emilio Cancio-Bello, Dave Prudhomme, and Mercedes Fuentes who not only did a lot of the benchwork but were always around to give a helping hand when things went wrong.

A special thanks must go to the people of Coulter Electronics, Incorporated. It was my initial contact with Dale Christenson and "Doc" Jones that inspired my interest in flow cytometry. My interactions with Dr. Al Brunsting were invaluable for the final formulation of parametric and nonparametric analysis. Bob Auer was essential in keeping our operation running smoothly and when all is said and done, it was the generosity of Dr. Wallace Coulter that made the project possible.

I especially wish to thank Dr. George Irvin for being my mentor, my employer, my physician, and most importantly, my friend.

Most of all I wish to thank my father for teaching me to have high goals; my mother for showing me how to work to obtain them; and finally my wife, Sandy, for making it all worthwhile.

TABLE OF CONTENTS

Page
LIST OF TABLES viii
LIST OF FIGURES ix

1. Introduction 1
2. History of flow cytometry 3

3. Fluorescent cytochemistry

3.1 Fluorescence 8
3.2 Cytochemistry 12

4. Explanation of DNA histograms

4.1 The cell cycle 20
4.2 DNA content vs cell age 24
4.3 Number of cells vs cell age 29
4.4 "Ideal" DNA histogram 40
4.5 "Real" DNA histogram 44

5. Information theory

5.1 Introduction 52
5.2 Flow cytometric structural information content.. 68

5.3 Information density

5.31 Line spread 94
5.32 Information density of a single message. 102
5.33 Information density of multiple messages 105

6. Parametric analysis I

6.1 Introduction 109

6.2 Smoothing

6.21 Introduction 110
6.22 Mathematics 113
6.23 Variable window size and optimization... 115

6.3 Capping

6.31 Introduction 127
6.32 Mathematics 128

6.4 Stripping

6.41 Introduction 131
6.42 Mathematics 132

6.5 Pattern recognition adjustment

6.51 Introduction 135
6.52 Mathematics 138

6.6 Translocation

6.61 Introduction 149
6.62 Mathematics 149

6.7 Doublet correction

6.71 Introduction 151
6.72 Mathematics 155

7. Parametric Analysis II

7.1 Introduction 174

7.2 Nonlinear least squares analysis

7.21 Introduction 175
7.22 Linearization 178
7.23 Steepest gradient 185

7.24 Marquardt compromise.................... 188
7.3 Normal vs lognormal noise distribution.......... 191
7.4 Standard deviation vs channel number............ 192
7.5 DNA/stain content vs fluorescent intensity...... 194
7.6 DNA histogram kernel functions
7.61 Introduction............................ 203
7.62 Multiple rectangle technique............ 217
7.621 Optimal number of S phase compartments.................... 226
7.622 Optimal size of S phase compartments.................... 228
7.623 Rate and transit time calculations.................... 229
7.624 Comparison with other approaches 235
8. Nonparametric analysis
8.1 Introduction.................................... 249
8.2 Smoothing and translocation.................... 251
8.3 Normalize to constant area..................... 252
8.4 Mean and standard deviation.................... 253
8.5 Statistical test................................ 254
8.6 Descriptor selection and use of Bayes' theorem.. 254
8.7 Discussion..................................... 255
9. Parathyroids
9.1 Anatomy.. 274
9.2 Physiology..................................... 276
9.3 Primary Hyperparathyroidism
9.31 Pathology.............................. 277
9.32 Flow cytometric detection of parathyroid hyperplasia.
9.321 Introduction.................... 279
9.322 Material and Methods............ 280
9.323 Results......................... 283
9.324 Discussion...................... 285
10. References.. 304

LIST OF TABLES

Table 4-1: Cell age vs clock time matrix. 31
Table 5.1-1: All combinations of four numbers summing to five.. 57
Table 5.32-1: Information gain vs CV and counts per channel. 108
Table 7.2-1: Linearization technique for solving nonlinear problems.. 180
Table 7.624-1: DNA histogram analysis comparison for test histogram 1.. 140
Table 7.624-2: DNA histogram analysis comparison for test histogram 2.. 141
Table 7.624-3: DNA histogram analysis comparison for test histogram 3.. 142
Table 7.624-4: DNA histogram analysis comparison for test histogram 4.. 143
Table 8-1: Classification of each histogram in the training set.. 272
Table 9-1: Parathyroid number distribution.................. 274
Table 9-2: Relative frequency of entities causing hyperparathyroidism.................................. 278

LIST OF FIGURES

Figure 2-1: Exploded view of TPS-1.......................... 6
Figure 3.1-1: Energy levels for mercury..................... 10
Figure 3.2-1: Structure of the groove binding probe, Hoechst 33258.. 16
Figure 3.2-2: Structure of the intercalating probe, PI...... 18
Figure 4.1-1: Scheme of the classical cell cycle............ 22
Figure 4.2-1: DNA content vs cell age....................... 25
Figure 4.2-2: Effect of cell heterogeneity on the shape of S 27
Figure 4.3-1: Number of cells vs cell age................... 33
Figure 4.3-2: Computer model of synchronous cells becoming asynchronous................................... 35
Figure 4.4-1: "Ideal" DNA histogram......................... 42
Figure 4.5-1: Ideal relation between SD and DNA content..... 48
Figure 4.5-2: "Real" DNA histogram and its synthesis........ 50
Figure 5.1-1: DNA histogram transmission process............ 62
Figure 5.1-2: Optimal yes/no question structure for finding a ball in four equally probable boxes................ 64
Figure 5.1-3: Optimal yes/no question structure for finding a ball in unequally probable boxes: 1/6, 1/6, 1/6, and 1/2... 66
Figure 5.2-1: Random signal with its Gaussian probability function for a 128 channel, 50,000 accumulation...... 78
Figure 5.2-2: Information analysis of a sine wave signal with an amplitude of 300............................. 80
Figure 5.2-3: Information analysis of a sine wave signal with an amplitude of 75.............................. 82
Figure 5.2-4: Information analysis of a sine wave signal with an amplitude of 30.............................. 84
Figure 5.2-5: Summary of information content vs sine wave amplitude... 86
Figure 5.2-6: Information analysis of a DNA histogram....... 88
Figure 5.2-7: Four DNA histograms with %CV's of 1, 3, 5, and 7... 90
Figure 5.2-8: Information content vs %CV................... 92
Figure 5.31-1: Fundamental model of line spread of a single message.. 96
Figure 5.31-2: Comparison of observed frequency histogram, binomial distribution, and Gaussian distribution..... 98
Figure 5.31-3: Linear line spread with multiple messages.... 100
Figure 6.2-1: Noiseless seventy Gaussian test case for smoothing optimization............................ 119
Figure 6.2-2: Noiseless test array [N-]..................... 121
Figure 6.2-3: Optimization [DIFF] array..................... 123
Figure 6.2-4: Smoothed and raw log DNA histogram............ 125
Figure 6.5-1: DNA histogram and its components with a CV of 6%.. 140
Figure 6.5-2: DNA histogram and its components with a CV of 3%.. 142
Figure 6.5-3: Family of error correcting curves for peak

height 144
Figure 6.5-4A,B: Effect of pattern recognition on estimation of percent S 146
Figure 6.7-1: DNA histogram of resting parathyroid cells.... 156
Figure 6.7-2: Semilog of Figure 6.7-1 158
Figure 6.7-3: Log aggregation transform of peak heights..... 160
Figure 6.7-4: Power aggregation transform of peak heights... 162
Figure 6.7-5: Log aggregation transform of entire histogram. 164
Figure 6.7-6: Power aggregation transform of entire histogram 166
Figure 6.7-7: Log aggregation transform of lymph node histogram 168
Figure 6.7-8: Log aggregation transform of ventral prostate histogram 170
Figure 6.7-9: Geometric relationships between transformed variables 172
Figure 7.5-1: Simplified model of cellular absorbance....... 197
Figure 7.5-2: Simplified model of cellular fluorescence..... 199
Figure 7.5-3: Fluorescence vs stain content 201
Figure 7.6-1: Dean and Jett kernel function 207
Figure 7.6-2: Fried kernel function 209
Figure 7.6-3: Fundamental limitation of the Fried technique. 211
Figure 7.6-4: Jett modified Fried kernel function 213
Figure 7.6-5: Multiple rectangle kernel function 215
Figure 7.6-6: Steps in the convolution of a rectangle with a Gaussian 222
Figure 7.6-7: Convolved rectangle 224
Figure 7.6-8: Schematic of multirectangle rate calculations 233
Figure 7.624-1: Synthesized histograms for the comparison of the multirectangle, multigaussian, and 2nd degree polynomial techniques 238
Figure 7.624-2: Multiple rectangle graphical analysis results test histogram 2 244
Figure 7.624-3: Multigaussian graphical analysis results of test histogram 2 246
Figure 7.624-4: Polynomial graphical analysis results of test histogram 2 248
Figure 8-1: Raw DNA histogram 258
Figure 8-2: Smoothed DNA histogram 260
Figure 8-3: Translocated DNA histogram 262
Figure 8-4: Two sets of processed DNA histograms 264
Figure 8-5: Mean histogram of type A and type B 266
Figure 8-6: Area of statistical difference between two sets. 268
Figure 8-7: Probability density functions for the descriptor 270
Figure 9.3-1: Typical DNA histogram from a normal parathyroid 290
Figure 9.3-2: Typical DNA histogram from an adenomatous parathyroid 292
Figure 9.3-3: Comparison of averaged normal and adenomatous parathyroid histograms 294
Figure 9.3-4: Hyperplastic index for each normal, "normal",

hyperplastic, and adenomatous hyperplastic indices... 296
Figure 9.3-5: Probability density functions for normal and adenomatous hyperplastic indices...................... 298
Figure 9.3-6: The calculated percent chance of being an adenoma for each normal, "normal", hyperplastic, and adenomatous parathyroid sample........................ 300
Figure 9.3-7: Parathyroid classification summary............ 302

The beauty of science is not in its complexity
but in its continuity.

1. Introduction.

Flow cytometry is a rapidly expanding technique where characteristics of cells flowing in a sheath of fluid can be acquired, stored, and analyzed. In Chapter 2 the history of flow cytometers and its operating characteristics are briefly outlined. The most common type of output from a flow cytometer is the DNA histogram; therefore, they are exclusively used in this study. In order to obtain a DNA histogram, cells must be prepared and stained with a suitable dye. The theory and practicalities associated with this fluorescent cytochemistry is discussed in Chapter 3. DNA histograms are a unique synthesis of some fundamental cell kinetic relationships. In Chapter 4 these principles and assumptions are discussed in depth. The actual amount of information contained in a DNA histogram (in BITS) can be calculated by using the principles of information theory. The equations developed in Chapter 5 have a number of interesting spin-offs from the engineering of optimal flow cytometers to the development of an optimal smoothing routine. The actual methods by which this information is extracted from DNA histograms is the subject of Chapters 6 and 7. These methods are under the heading of parametric analysis because they attempt to reduce histogram data to a small number of biologically meaningful parameters. The most sophisticated analysis discussed is parametric analysis II or the multiple rectangle technique. This method is compared with other popular techniques in Chapter 7. In Chapter 8, the statistical technique of nonparametric analysis is introduced and is used in the clinical problem of identifying parathyroid hyperplasia in Chapter 9.

It was my ultimate objective in this project to play a part in the establishment of flow cytometry as a practical and clinical tool. Flow cytometry holds great promise in helping medicine along its long and arduous path from being an art to becoming a science.

2. History of flow cytometry (1).

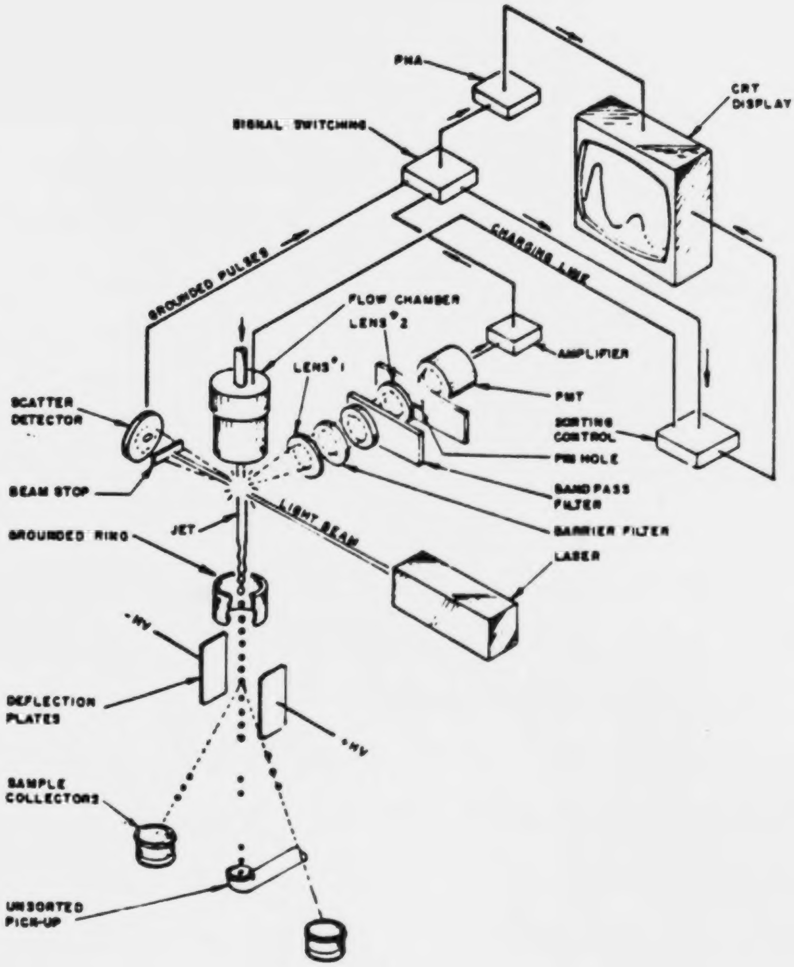
In the early 1950's Wallace Coulter constructed an apparatus in a friend's basement that would automatically count cells. The counter was based on a deceptively simple principle. Cells, as they pass through a narrow orifice, displace the surrounding medium. If the medium has a different resistivity than the cells, the total resistance through the orifice will change. This change in resistance is very nearly proportional to the volume of the cell. If one imposes a constant current through the orifice, then this change in resistance causes a proportional change in voltage. This voltage can easily be amplified and displayed as a pulse on an oscilloscope or a count on a multichannel analyzer. One can then design the electronics such that counts of particular voltages/cell sizes can be accumulated. By forcing a known volume of media through the orifice, the actual concentration of a particular cell population can be obtained. This device truly marks the beginning of automated cytometry. This machine is known all over the world as the Coulter Counter and its impact on medicine can be appreciated by the simple fact that over 80% of the hospitals in the country have at least one.

In 1965 Mack Fulwyler took this principle one step further. He constructed a special nozzle that would allow cells to enter into the core of a fast moving thin sheath of fluid. As the cells passed through the orifice, they caused a voltage fluctuation that was proportional to the cell's volume. By vibrating the nozzle at high frequencies with a piezoelectric crystal, perturbations were introduced into the sheath which eventually resulted in the sheath breaking up into minute droplets further down stream. The circuitry was timed such that, a cell of a particular size was sensed and then at the precise moment (just before the drop broke away from the stream) the drop containing the cell was charged. The charged drop could then be deflected either to the right or left as it passed between two deflection plates. This was the first cell sorter and its full impact on medicine is still to be made.

In the early 1970's (2) a team effort at a division of Coulter Electronics called Particle Technology, Inc (P.T.I.) made the first Two Parameter Sorter (TPS-1) under the direction of R. Auer (project engineer) and M. Fulwyler (president). An exploded view of the TPS-1 is shown in Figure 2-1. Instead of using cell volume as a measurement parameter, this machine used an argon laser to illuminate the cells as they passed through the beam. Two different kinds of information can be obtained from cells in this configuration: light scatter and fluorescence. As the cell enters the coherent light source, it scatters light in a complex way. The mathematics that describe this scatter signal are highly involved and are described more fully elsewhere (3). If only low scatter angles are allowed to be integrated, this signal is usually related to cell volume. At 90° from the laser-sheath plane is a lens, filter, and photomultiplier tube (PMT). This system captures the emitted fluorescence of cells as they pass through the laser beam. If cells are tagged with a stain that fluoresces at the laser wavelength, their fluorescence can be detected, amplified, and stored on a multichannel analyzer as a fluorescence histogram. The number of possible applications of this device is only limited by the number of suitable stains and one's imagination.

Since this device involves the flow and measurement of cells, it is called a flow cytometer. Improvements and extensions of flow cytometers are in progress all over the world: L. Wheeless and P. Horan at the University of Rochester; R. Auer and A. Brunsting at Coulter Electronics; P. Mullaney and J. Jett at Los Alamos Scientific Laboratory; M. L. Mendelson and M. A. Van Dilla Laurence Livermore; L. Herzenberg at Stanford Medical School; M. Melamed at Sloan-Kettering Cancer Center; D. Arndt-Jovin and T. Jovin at Max Planck Institute; R. Leif at the Papanicolaou Cancer Research Institute, C. J. Herman at the National Cancer Institute; and M. Ingram and B. Price at the Institute of Cell Analysis.

Figure 2-1: Exploded view of TPS-1.


PHA
CRT DISPLAY
SIGNAL SWITCHING
GROUNDED PULSES
CHARGING LINE
FLOW CHAMBER
LENS #2
LENS #1
AMPLIFIER
PMT
SCATTER DETECTOR
SORTING CONTROL
PIN HOLE
BEAM STOP
BAND PASS FILTER
BARRIER FILTER
JET
LIGHT BEAM
GROUNDED RING
LASER
-HV
DEFLECTION PLATES
+HV
SAMPLE COLLECTORS
UNSORTED PICK-UP

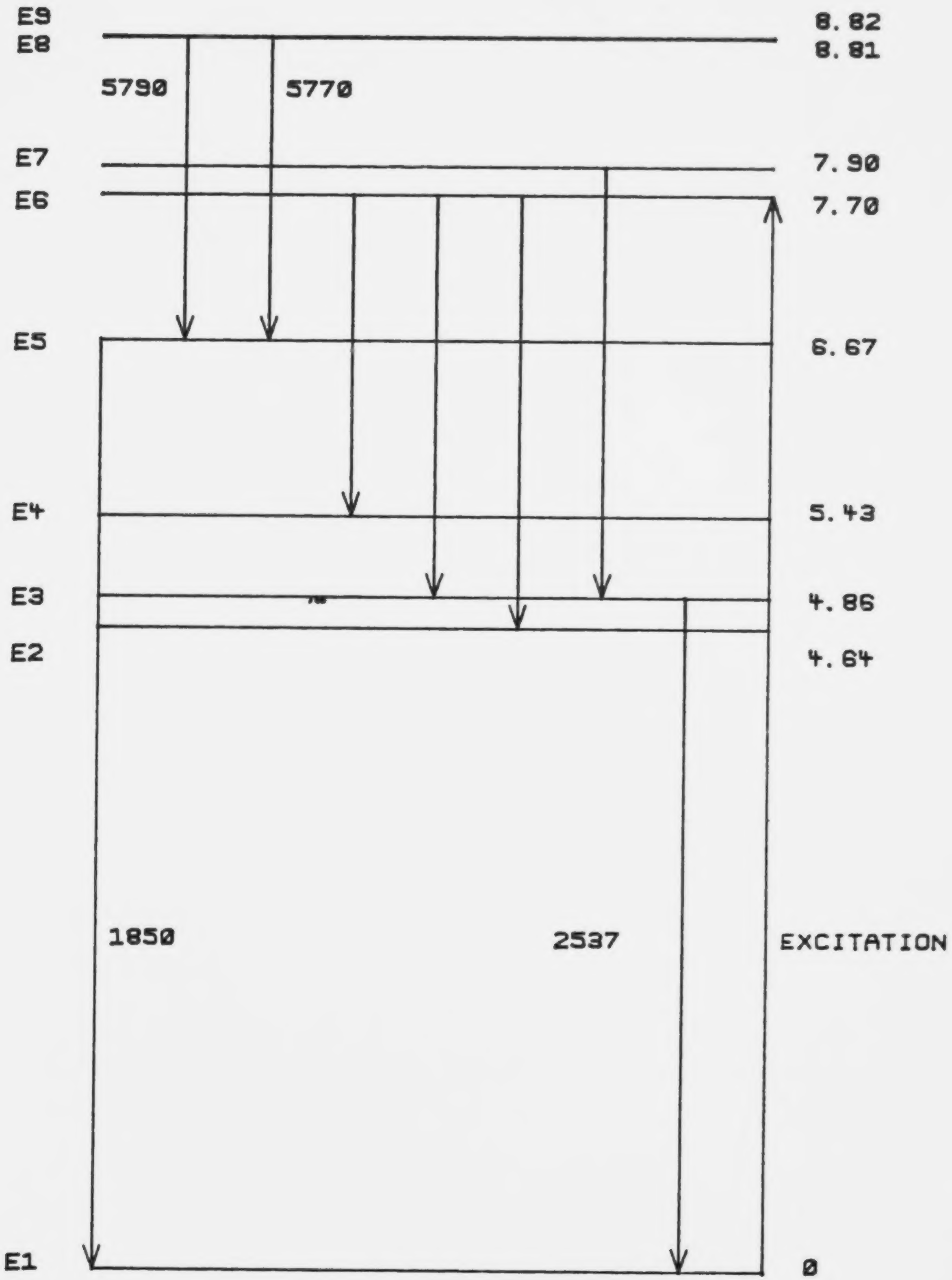
3. Fluorescent cytochemistry.

3.1 Fluorescence.

In order to understand fluorescence we must first understand the relation between absorption and emission (4). At ordinary temperatures an atom spends almost all of its life in its lowest energy state (E1, see Figure 3.1-1). If by chance an incoming photon gives up its energy to an atom, the atom can be momentarily raised to one of its excited states. An atom usually stays in an excited state for only about 10^{-8}s, and then if left to itself emits a photon of longer wavelength and spontaneously "jumps" to a lower state. After a very short time, the atom has been "shaken down" to its normal state, and one or more photons have been emitted in the process. The emission of light which is thus dependent on the absorption of other light is called fluorescence. We can see from Figure 3.1-1 that fluorescent radiation always has a longer wavelength than the excitation radiation. This phenomenon is called the "Stokes shift". The ratio between the number of fluorescent photons and the incident or absorbed photons is called the quantum yield or quantum efficiency (5). Compounds that have zero quantum yield are not fluorescent. Fluorescence is usually a more sensitive measurement than absorbance. The fundamental difference between these two measurements can best be explained by example (6).

If one were to turn out one light in a sunlit room, the change in brightness would seem a lot less than if one lit a match in a completely dark room. This is because the signal to background ratio for the match in a dark room is much higher than the light change in a sunlit room. Absorbance measurements are made against a bright backgound whereas fluorescent measurements are made against a dark background. The high sensitivity of fluorescence is the major advantage over absorbance; however, it is not without disadvantages (Chapter 7.5). The method by which one ensures a dark background for fluorescent measurements is to interpose a filter between the detector and the sample that absorbs the exciting wavelength of light and transmits the fluorescent wavelengths. A filter of this type is called a "high-pass" filter.

Figure 3.1-1: Energy levels for mercury. The upward arrow signifies the excitation of the electron to a higher and more unstable energy level. Since the electron descends toward its stable conformation in small steps, the energy released are photons with wavelengths longer than the exciting wavelength. The five digit numbers are the wavelength in Angstroms and the three digit numbers along the right side are energy levels in eV.


E9
E8
8.82
8.81
5790
5770
E7
7.90
E6
7.70
E5
6.67
E4
5.43
E3
4.86
E2
4.64
1850
2537
EXCITATION
E1
0

3.2 Cytochemistry (6).

In order to obtain specific information about a cell in a flow cytometer, one generally has to stain the desired cellular component with a fluorescent compound. These compounds are commonly referred to as either chromophores or probes. There are two major classes of chromophores: covalent binding and noncovalent binding. The classical example of a covalently bound probe is pararosaniline in the feulgen DNA staining technique. This procedure involves a fixation of cells, treatment with HCl which releases purines thereby exposing aldehyde moities, and a covalent binding of pararosaniline to these aldehyde groups. The cells are then washed and can be run immediately or stored for a considerable period of time. There are two classes of noncovalent binding probes: groove binding and intercalating. The nature of the binding force is hydrophobic and/or electrostatic and/or VanderWaals. Groove binding probes have a long extended structure (see Figure 3.2-1) and either bind to the major or minor groove of DNA. Some examples of groove binding probes are the Hoechst dyes, mithromycin, chromomycin, and 7-aminoactinomycin. Intercalating DNA probes actually stack between the base pairs and extend and rotate the DNA structure. Most of these molecules are 3 ringed and planar (see Figure 3.2-2).

Some examples of intercalating probes are propidium iodide (PI), ethidium bromide (EB), acridine orange (AO), and the acriflavins.

There are a number of characteristics of probes that are to be considered for their selection (6).

1. Specificity. Does the compound only attach to the target component or does it attach to other cellular components as well? For example, the intercalating stain, PI, not only will attach to double stranded DNA but also double stranded RNA.

2. Stoichiometry. If there is twice as much of the target component, will there be twice as much staining? Stains should be in saturating quantities to ensure stoichiometry. However, as the concentration of the stain increases, other problems arise (see Chapter 7.5). An example of a stain that is very specific but lacks stoichiometry is 7-aminoactinomycin. This stain is specific for GC base pairs; however, there is a presumable interference between it and histones that make it not stoichiometric.

3. Absorbance and fluorescence spectrum matching the flow cytometer. Most of the flow cytometers involve

lasers for excitation and thus excite at only a few wavelengths. A probe that is excited only in ultraviolet regions, e.g. Hoechst, would be unsuitable for flow cytometers that only have a 488nm excitation line (e.g. low power Argon lasers).

4. Stain stability. Some chromophores (eg. AO) are extremely unstable and must be analyzed immediately after staining. If samples are to be analyzed greater than 15 minutes after staining, this probe would be an unsuitable choice.

5. Changing morphology. If cells are to be examined after staining, then neither the probe nor the procedure should alter the cell structure. Mithromycin is a good stain in this respect because of its mild staining conditions.

6. Vital stain. Some stains will stain only live cells and some stains will stain only dead cells. This characteristic can be an advantage as in the case of live-dead differentiation using differential PI concentrations (7).

7. Fixation. If one's application involves the storing

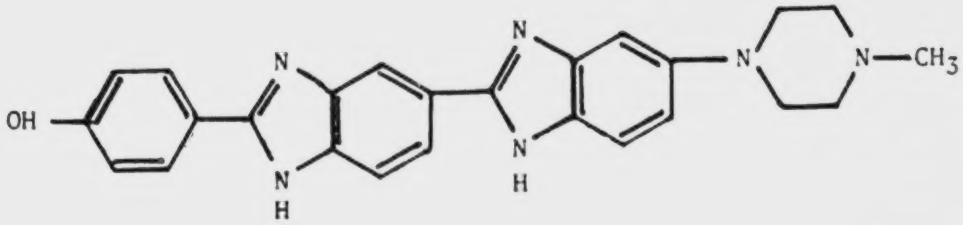
of samples, then the stain must be suitable for fixation. There have been a number of fixation techniques proposed for PI (8,9).

8. Time. Procedures that require excessive manipulations are not as popular as one-step procedures such as the one proposed by A. Krishan (10).

9. Quantum yield. A useful feature of some stains is a quantum yield increase upon attachment to the target component. If the increase is substantial, then there are fewer protems in background fluorescence and washing procedures.

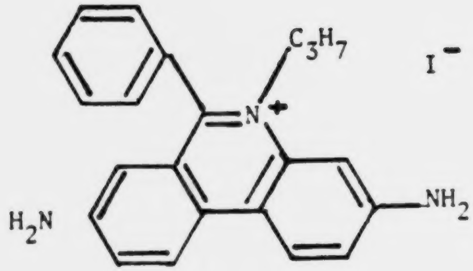
The staining technique that we use is a slight modification of the low salt technique proposed by Vindelov (11). This procedure involves subjecting cells to hypotonic conditions which, in addition to detergent, causes complete cytoplasmic dissolution. The nuclei are then stained with propidium iodide. Large quantities of RNAse are present to degrade any double stranded RNA. The details of the staining procedure for parathyroid tissue can be found in the material and methods section of Chapter 9.322. The nuclei are filtered and run through the flow cytometer to yield a DNA histogram. The theoretical interpretation of a DNA histogram is the subject of the next chapter.

Figure 3.2-1: Structure of the groove binding probe, Hoechst 33258, The long extended nature of this molecule allows it to bind either to the major or minor grooves in DNA.


OH
N
N
H
N
N
H
N
N
CH3

HOECHST 33258

Figure 3.2-2: Structure of the intercalating probe, propidium iodide. The flat planar aspect of this molecule enables it to fit between the base pairs of DNA or double stranded RNA.


C_3H_7
I^-
N^+
H_2N
NH_2

PROPIDIUM IODIDE (PI)

4. Explanation of DNA histograms.

4.1 The cell cycle.

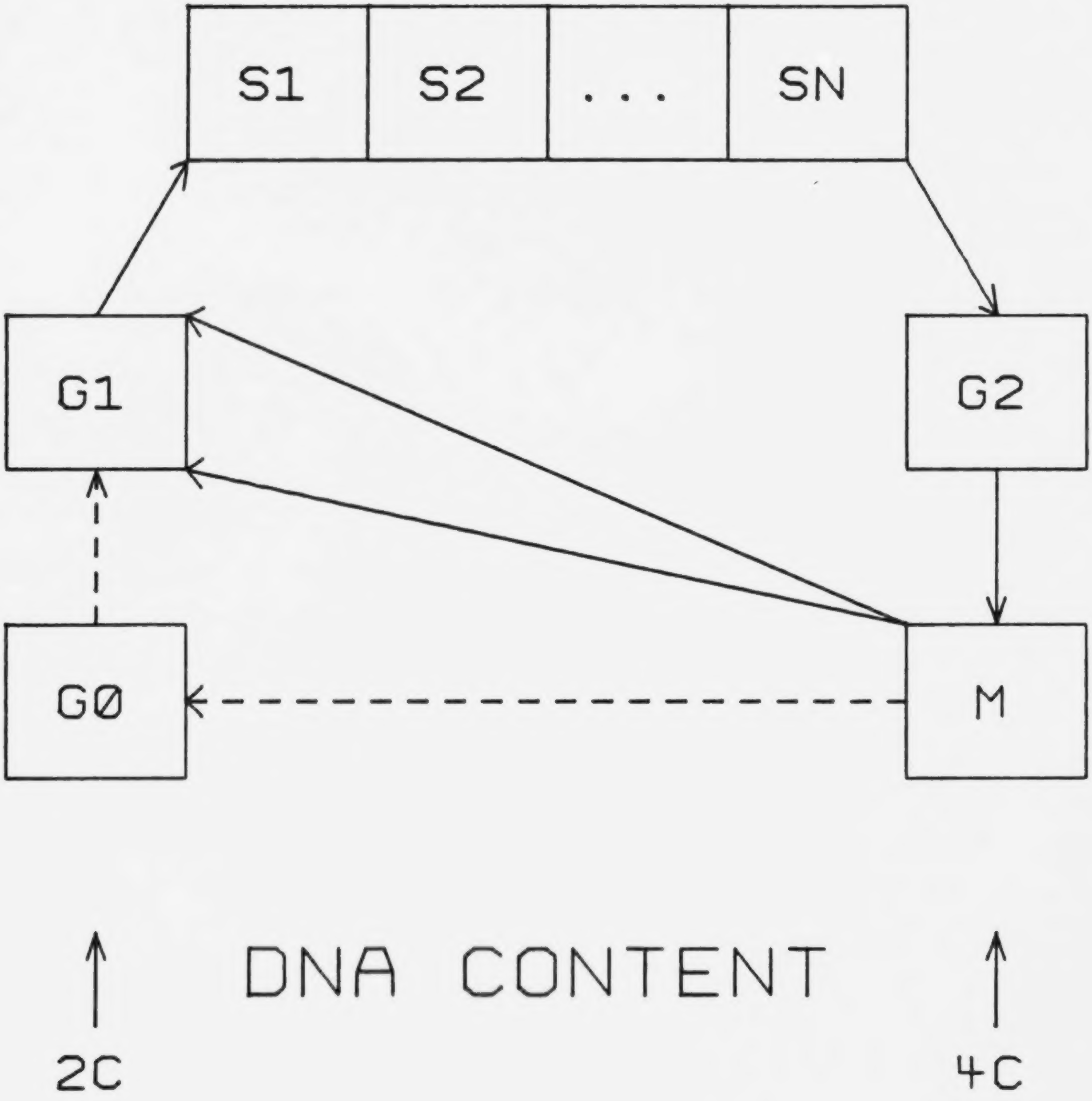
A histogram is a plot of frequency on the ordinate and the measured parameter on the abscissa. The measured parameter in this case is the content of DNA per cell nucleus. This function is really the result of two more fundamental relationships: DNA content versus cell age and number of cells versus cell age. In order to understand why these relationships have their characteristic appearance, we must have a firm concept of cell cycle dynamics. A scheme of the classical cell cycle is shown in Figure 4.1-1. The cell cycle compartments are drawn such that their horizontal position reflects their respective DNA content. Cells that contain only one complement of DNA from each parent (2C) are referred to as diploid cells. Cells that have duplicated their genome and thus have 4C amount of DNA are called tetraploid cells.

The cell cycle is classically divided into the following phases: G1, S, G2, M, and G0. The cell cycle phase of G1 was historically considered to be a time when diploid (2C) cells had little observable activity. Since this time preceded DNA synthesis, the term Gap 1 (G1) was coined. We now know that there is quite a bit of transcription and protein synthesis during this phase. The main enzymes that are produced are those involved in DNA synthesis, e.g. polymerases. At a certain point in the cell's life, the DNA synthetic machinery turns on. This phase of the cell's life is labeled S for Synthesis. As the cell proceeds through this phase, its DNA content increases from 2C to 4C. At the end of S, the cell has duplicated its genome and now is in the tetraploid state. After S phase, the cell again enters a phase that historically was thought to be quiescent. Since this phase is the 2nd Gap region, it is referred to as G2. In the G2 phase, the cell is producing the necessary proteins that will play a major role in cytokinesis. After a highly variable amount of time, the cell enters mitosis (M). The subphases of M (prophase, metaphase, anaphase, telophase) have been well described elsewhere (12). DNA content remains constant at 4C until the cell actually divides at the end of telophase.

Cell age is defined as beginning at the point immediately following daughter cell separation and ending just before separation. The length of time it takes a cell to traverse the cell cycle is called the cell cycle time (Tc). The length of time needed to traverse a cell cycle phase is called the transit time, e.g. G1 transit time (T_{G1}).

Figure 4.1-1: Scheme of classical cell cycle. The cell cycle compartments are drawn such that their horizontal position reflects their respective DNA contents. The resting G0 cells enter the cell cycle when stimulated. They begin in G1 where the cell prepares itself by making proteins necesary for DNA synthesis. During S phase the cell is duplicating its genome. When the cell has doubled its normal complement of DNA, it enters G2 where it begins making proteins necessary for cytokinesis. Mitosis (M) is the phase were the cell visibly condenses its chromosomes (prophase), lines them up along the equatorial plane (metaphase), separates the daughter chromosomes (anaphase), and then divides (telophase). The newly formed daughter cells have a choice of either re-entering the cycle at the level of G1 or leaving the cycle by entering G0.

CELL CYCLE


S1
S2
...
SN
G1
G2
G0
M
DNA CONTENT
2C
4C

4.2 DNA content vs cell age.

Figure 4.2-1 shows a plot of the average DNA content per cell versus cell age. Cell age is conventionally normalized such that Tc=1. During G1 the DNA content is 2C and invariant. During G2 and M the DNA content is 4C. The shape of DNA content vs cell age during S is a function of the average rate of DNA synthesis and the heterogeneity of the cell population. The faster the rate of DNA synthesis, the steeper the S phase curve is going to be. The converse is also obviously true. The influence of heterogeneity on the shape of S is more subtle and best shown by example (see Figure 4.2-2). Three representative populations with constant and equal rates of DNA synthesis are schematized in the upper aspect of this figure. If these populations have the same cell cycle time, but different G1 and G2 transit times, the average of the three populations is going to have a rounding effect on the final S phase. If the heterogeneity of the population were not appreciated, it would be tempting to believe that the average rate of DNA synthesis slowly peaked at some mid-S cell age value and then decreased. All DNA histogram rate calculations reported (13,14) have to tacitly assume a homogeneous population of cells with constant transit times because there is at present no known way to compensate for this effect.

Figure 4.2-1: DNA content vs cell age. During G1 the DNA remains constant at 2C DNA content. As DNA synthesis continues, the DNA content increases from 2C to 4C. The rate of increase is a function of the rate of DNA synthesis. In G2 and M the DNA content is constant at 4C. After the cell divides this process repeats itself if the cells remain in the cycle.

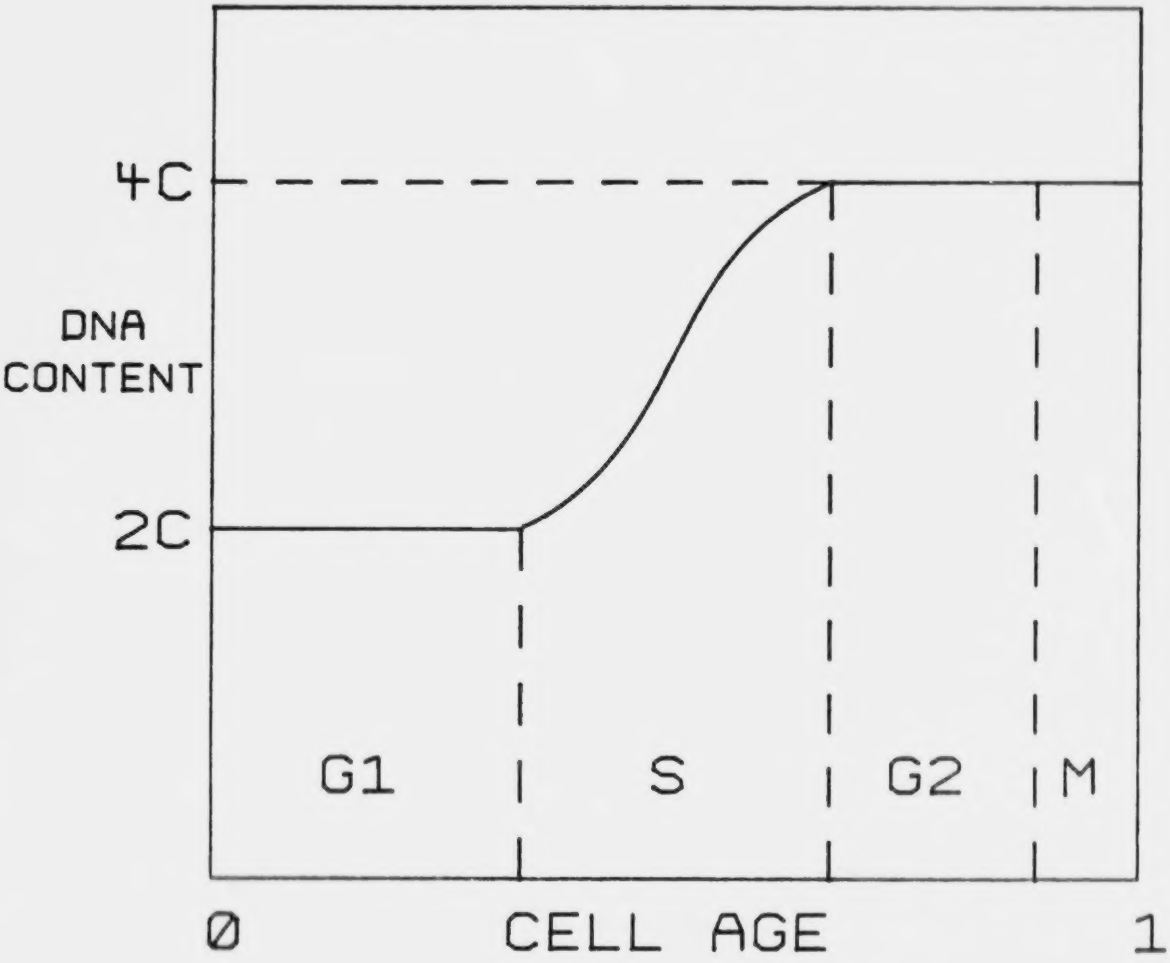
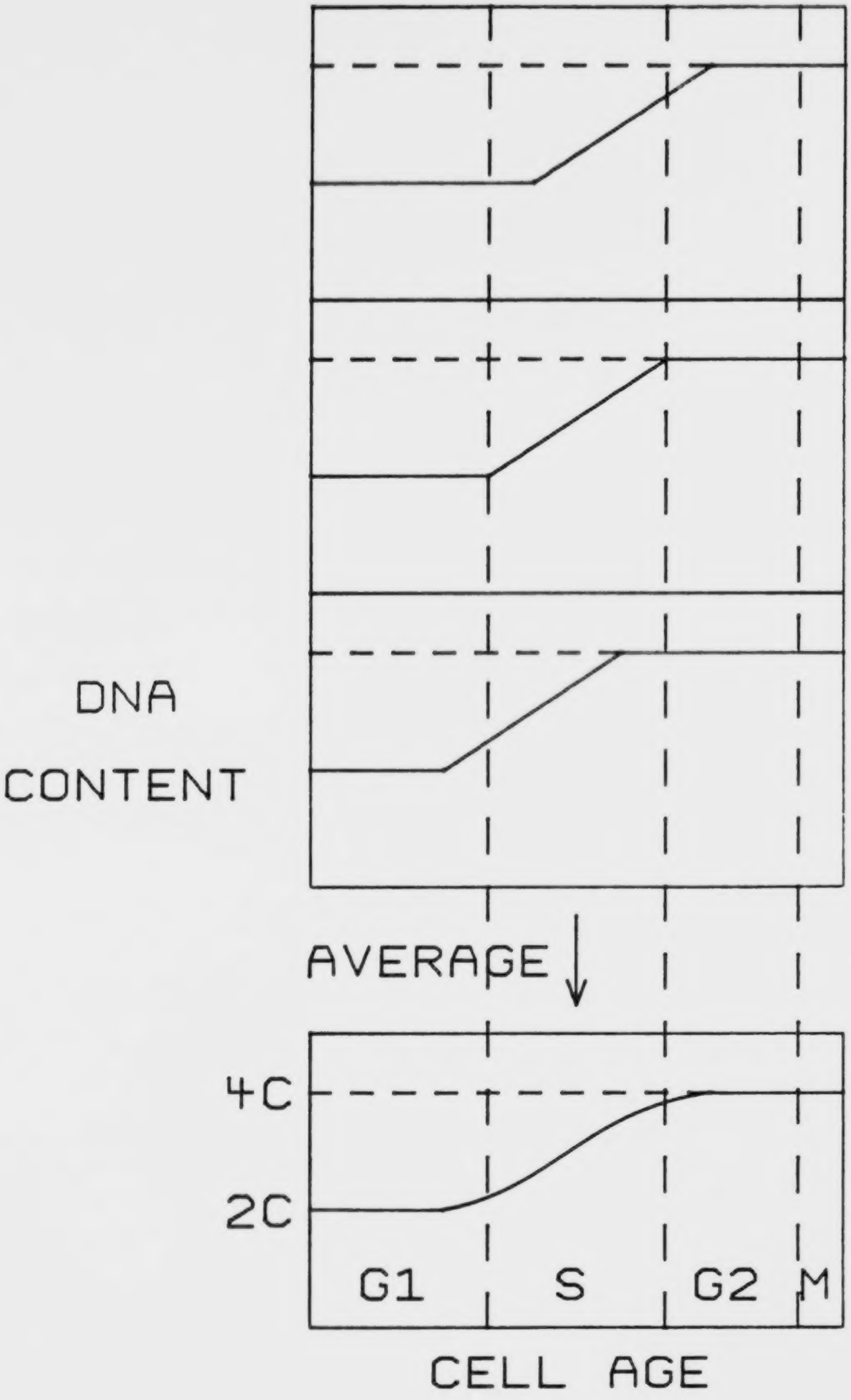

4C
DNA
CONTENT
2C
G1
S
G2
M
0
CELL AGE
1

Figure 4.2-2: Effect of cell heterogeneity on the shape of S phase. Cells with virtually identical rates of DNA synthesis, will give a deceiving picture of the average rate of DNA synthesis if the transit times vary from one cell to another.


DNA
CONTENT
AVERAGE
4C
2C
G1
S
G2
M
CELL AGE

4.3 Number of cells vs cell age.

The relationship between number of cells and cell age is simple and fundamental, but difficult to prove analytically. This relation can be appreciated better by constructing an analogous example. Let us say we have a running track of some arbitrary circumference and a few runners are randomly placed around the track. This track, instead of having a finishing line, has a doubling line. When a runner crosses this line another runner enters the race at that point. The runners are of varying ability; therefore, their speeds are randomly distributed. Now, let us start the race. The first racer that crosses the line is joined by another. As the race continues more and more people are on the track. It is intuitively obvious that there is going to be a large crowd directly after the doubling line and less people just before the doubling line. In fact, since the line is a doubling line, there should, on the average, be 2 times the number of people directly after the line as just before it. But what is the relationship between people density and track length between these two end points? Is it a straight line, exponential, hyperbola, or perhaps some more esoteric relationship? The answer turns out to be an exponential which is discussed more fully below. The parallelism between this example and cells cycling is almost perfect.

The length of the track is related to cell age and the density of runners is analogous to the density of cells. Figure 4.3-1 shows this exponential relationship between number of cells and cell age. Since cell age begins just after cell division (To=0), there are 2X number of cells and at the end of cell age (Tc=1), there are 1X cells.

Before attempting to describe this exponential relationship more analytically, let us test it empirically. If we construct a computer program that models the cells as they mature at different rates, we can let the computer run through a number of cell cycles and test the final result. Let us give our model the worst possible of starting conditions where all the cells are clustered together at the beginning of the cell cycle. Figures 4.3-2A,B,C,D show the distributions as a function of time. When cells are clustered together as in Figure 4.3-2A, they are called synchronous. When the cells finally distribute themselves in a time constant distribution such as shown in the last panel in Figure 4.3-2D, they are called asynchroucus. The function that best described the final shape of cell age vs cell number was the exponential with end points of 82.27 and 40.40 respectively.

Table 4.3-1: Cell age vs clock time matrix.

		Cell age ----½					
		1dt	2dt	...	Tc-dt	Tc	Total number
	dt	2dx	1dx	...	1dx	1dx	x+1dx
Clock	2dt	2dx	2dx	...	1dx	1dx	x+2dx
	.	.	.	...	.	.	.
Time	.	.	.	...	.	.	. Y=X*exp(T*ln2/Tc
	.	.	.	...	.	.	.
	Tc-dt	2dx	2dx	...	2dx	1dx	2x-dx
	Tc	2dx	2dx	...	2dx	2dx	2x
Age Distribution	-----½	2x	2x-dx	...	x+2dx	x+1dx	

Y = 2X*exp(-T*ln2/Tc)

Another way of appreciating this exponential relationship is by close examination of Table 4.3-1. This table depicts the relationships between number of cells, cell age, and clock time. Both clock time and cell age are broken into differential time units, dt. This matrix describes how a differential number of cells, dx, changes as a function of both of these time parameters. Along the right side of the table is the total number of cells at each differential clock time. As expected at the end of one cell cycle (Tc), the total number doubles. By summing the columns, one obtains the distribution of cells as a function of cell age. Notice the inverse symmetry between total number of cells and the age distribution. The formula that calculates the number of cells as a function of time is classically written as

$$Y = X * 2^{T/Tc} \quad (4.3\text{-}1)$$

Let us see if this equation makes sense at the end points. When T=0, Y=X and when T=Tc, Y=2X. This is what we would expect. Since the age distribution is inversely related to eq 4.3-1, its formula is

$$Y = 2X * (1/2)^{T/Tc} \quad (4.3\text{-}2)$$

At T=0, the beginning of cell age, Y=2X and at T=Tc, just before cell division, Y=X; thus, it satisfies our endpoint criterion. If the cell cycle time is arbitrarily set to unity (Tc=1), the formula transforms to

$$Y = 2X * (1/2)^{T} \quad (4.3\text{-}3)$$

Equation 4.3-3 can be rewritten in the base of e which is mathematically easier to manipulate.

$$Y = 2X * e^{-\ln 2 * T} \quad (4.3\text{-}4)$$

It should be noted at this point that equation 4.3-4 is valid only if the population is exponentially growing (no cell death) and is asynchronous. If either one of these conditions is not satisfied, this equation becomes considerably more complex.

Figure 4.3-1: Number of cells vs cell age. In a nonsynchronous, exponentially growing population of cells, there is an exponential relationship between number of cells and cell age.

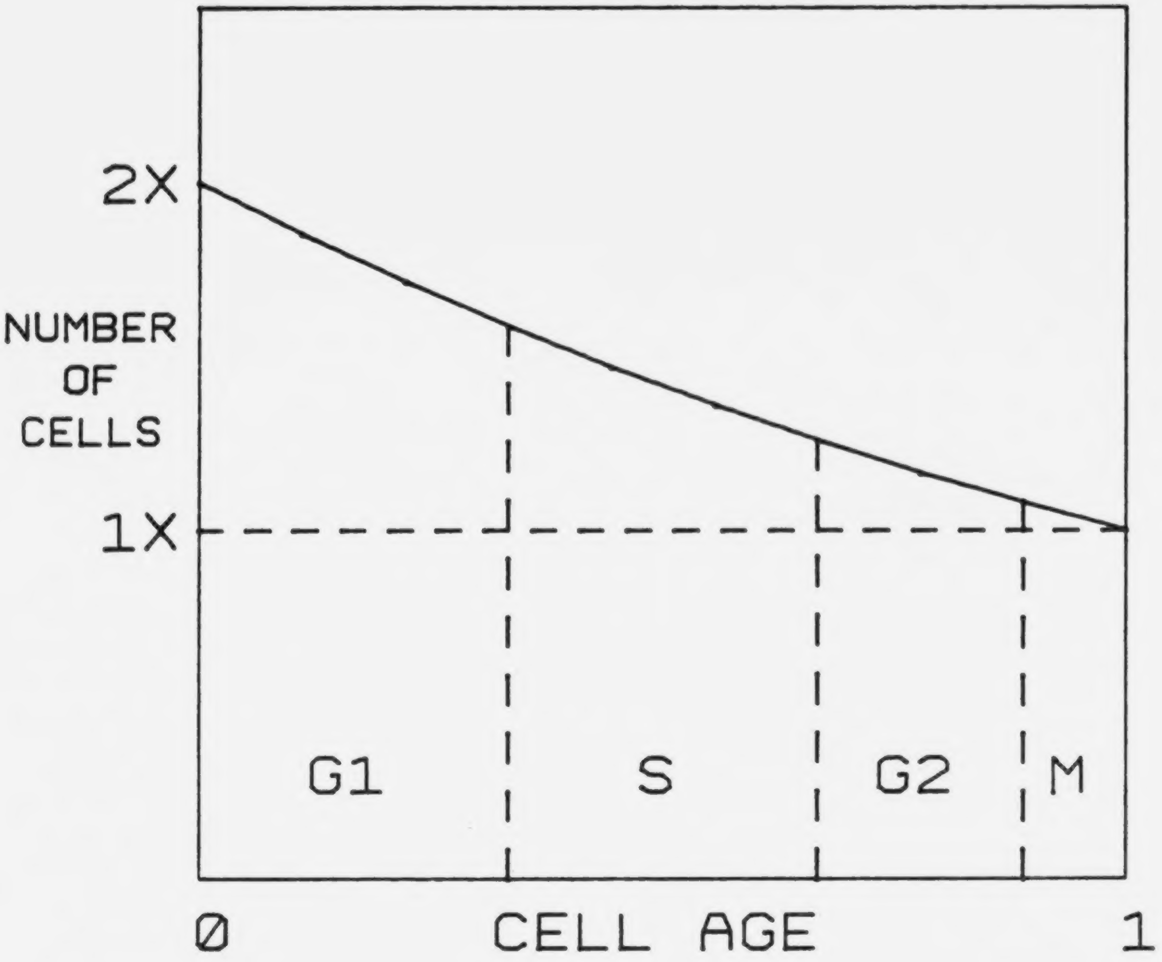
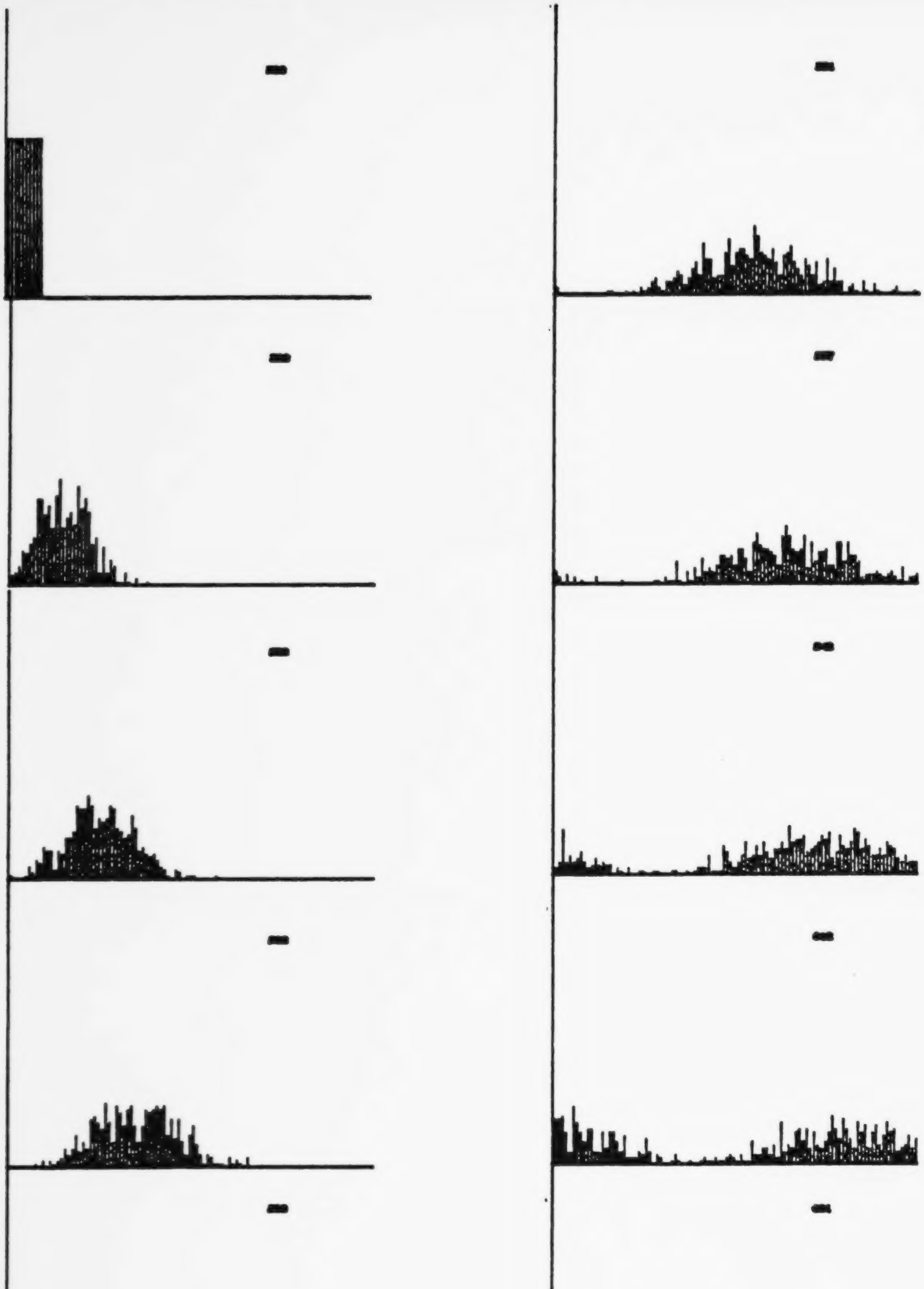
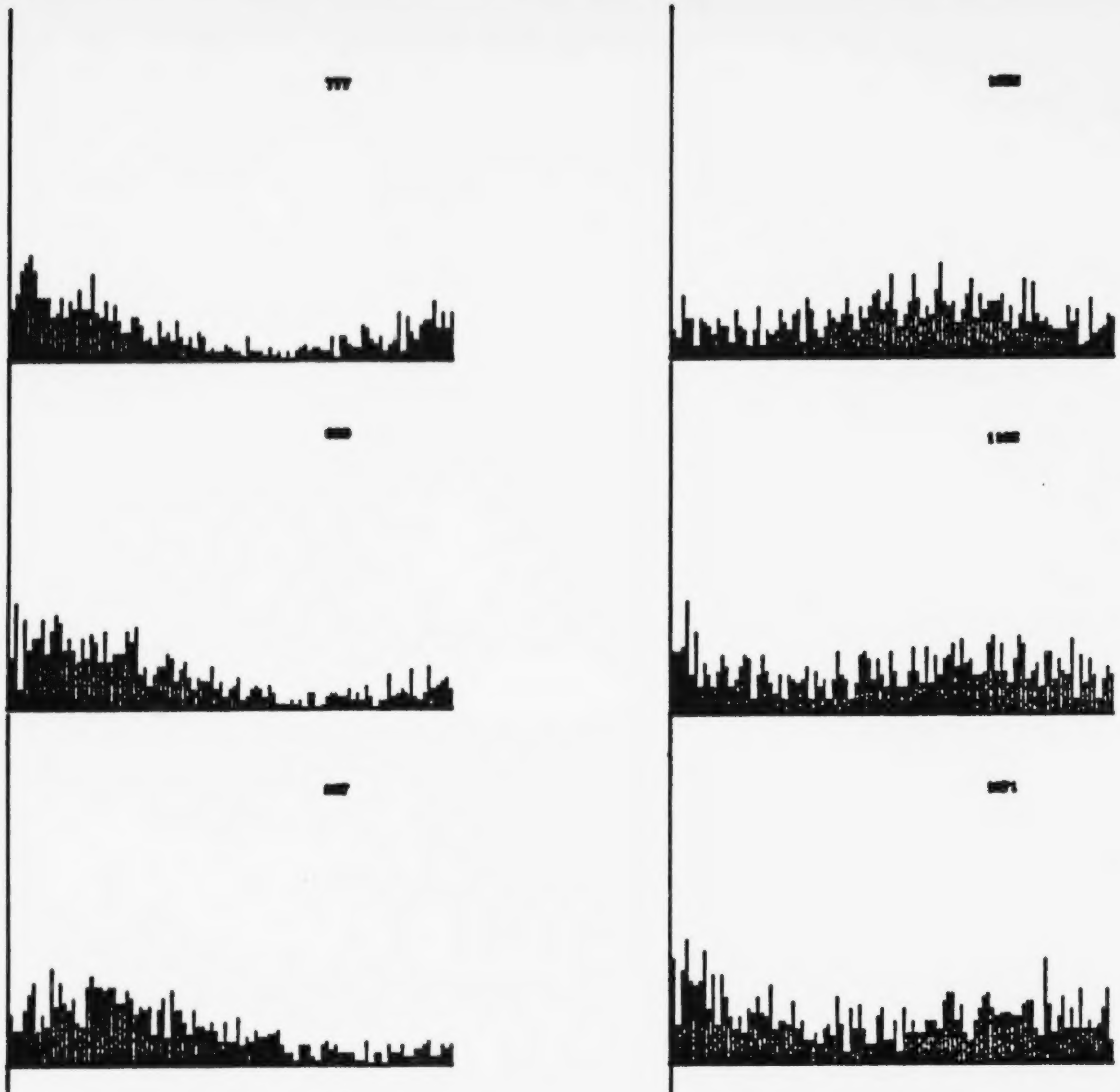
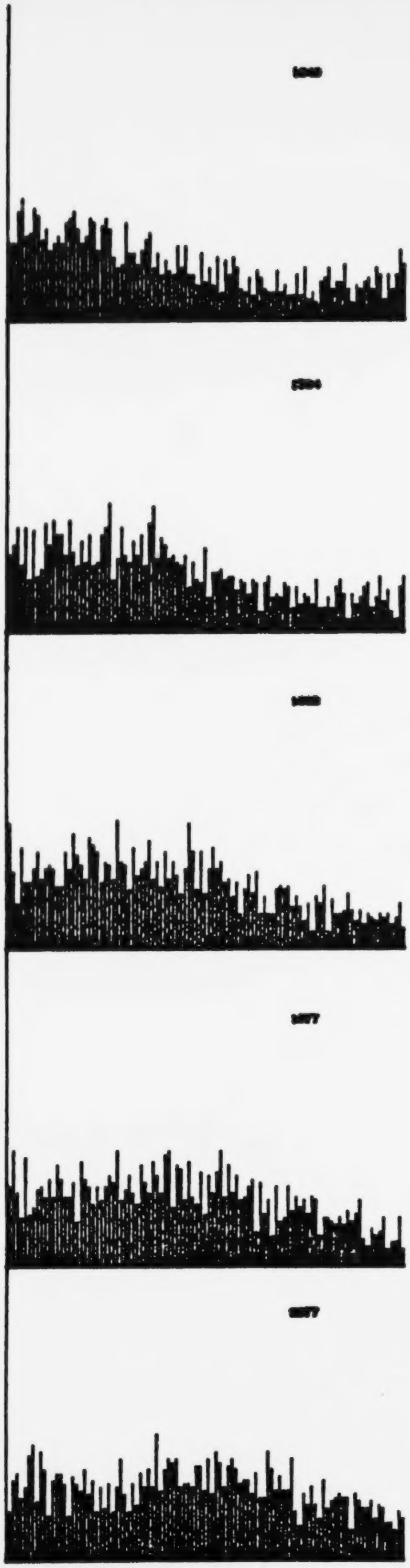
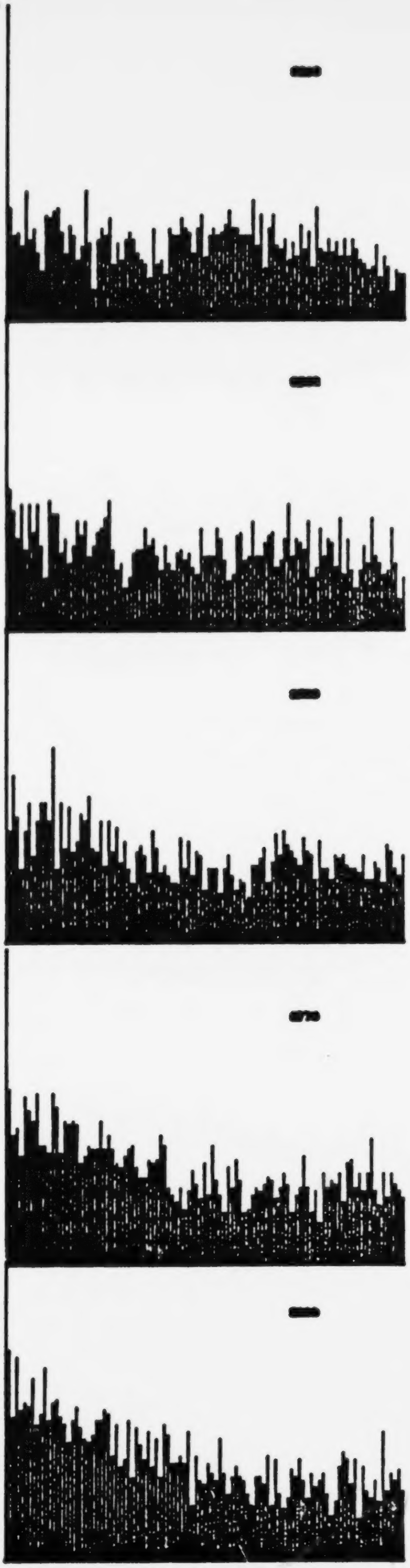

2X
NUMBER OF CELLS
1X
G1
S
G2
M
0
CELL AGE
1

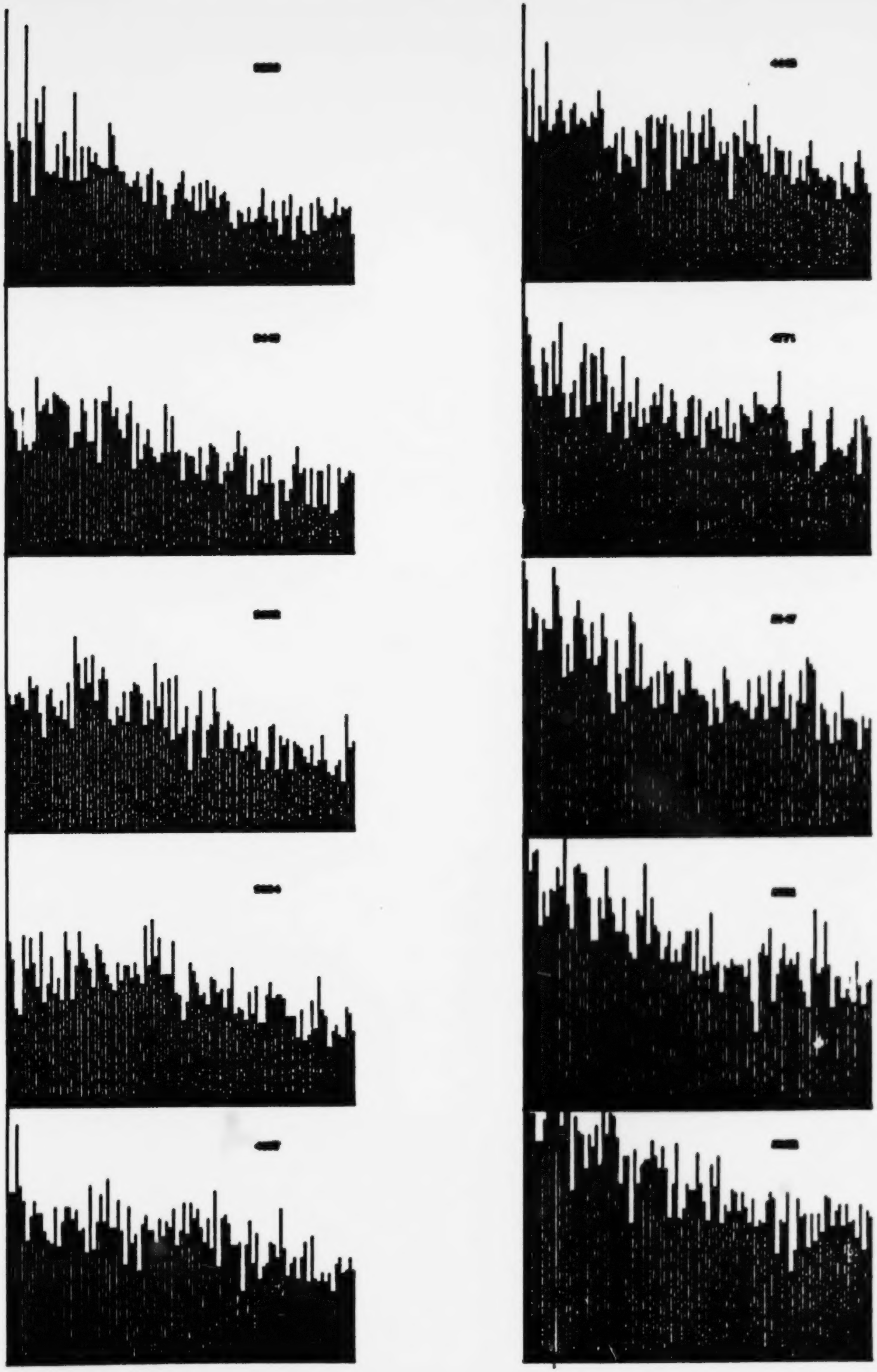
Figure 4.3-2A,B,C,D: Computer model of synchronous cells becoming asynchronous. Because cells have slightly different cell cycle times, a synchronous population will eventually become asynchronous if they are allowed to grow. This computer simulation shows a multitude of graphs of number of cells vs cell age at different times after the start of cell growth. The important aspect of this demonstration is the final exponential shape of those curves near the end when they totally lose their synchrony.









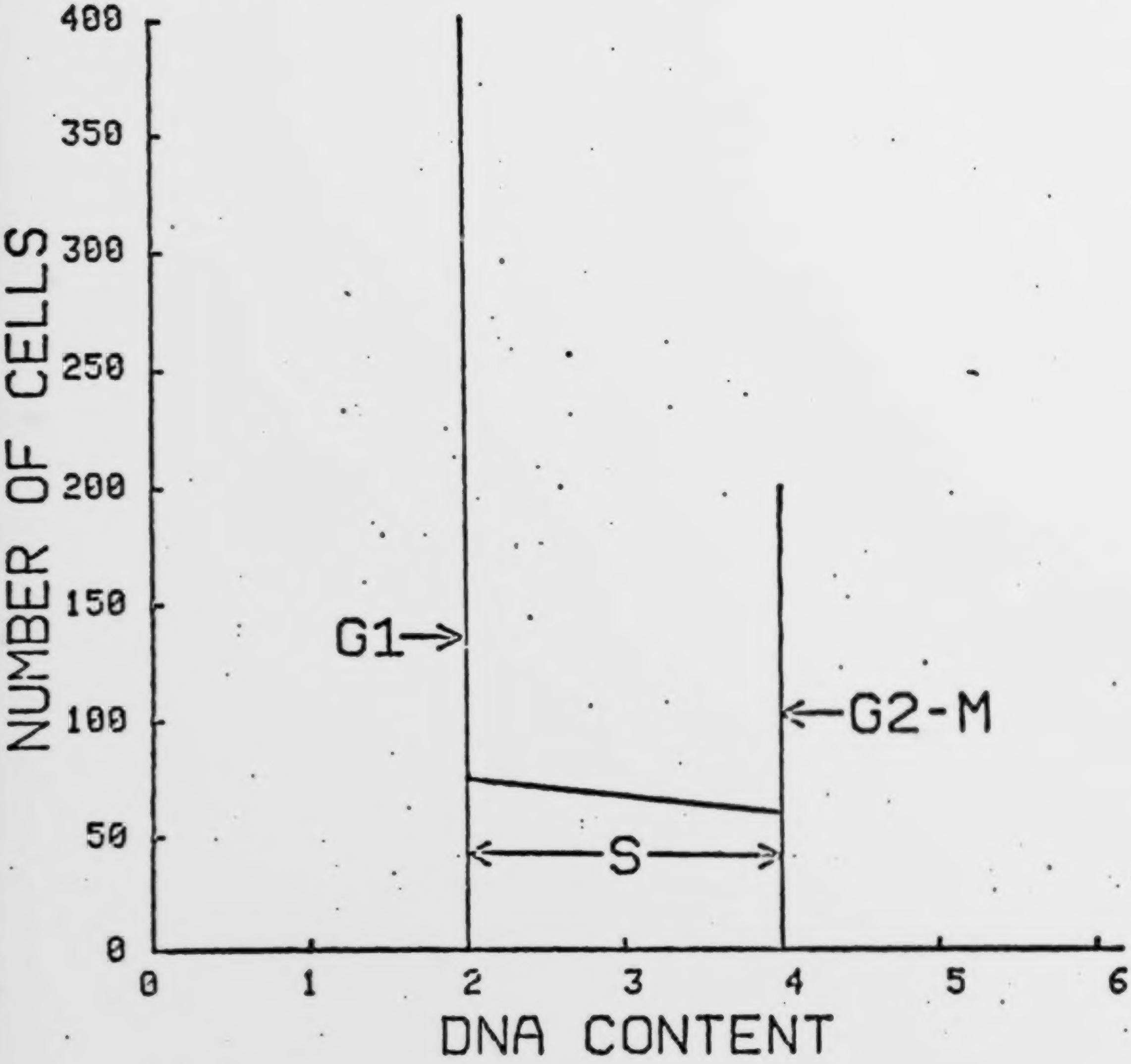


4.4 "Ideal" DNA histogram.

An "ideal" DNA histogram is the result of the combination of the two relationships described above: DNA content vs cell age and number of cells vs cell age. The common parameter of time (cell age) drops out and we are left with number of cells vs DNA content (see Figure 4.4-1). Since cells that are in G1 phase of the cell cycle have a constant DNA content (2C), they become a spike whose height equals the area under Figure 4.3-1 over the G1 transit time region. Since G0 cells also have 2C amount of DNA, they also contribute to the height of the spike. This spike is appropriately named the G0-G1 spike. The cells in G2-M have an invariant amount of 4C DNA and therefore can be represented as a G2-M spike whose height equals the number of cells in G2-M. The S phase is slightly more complicated. As the DNA content increases, we move from left to right on Figure 4.2-1. The faster the DNA content increases, the faster we move along the DNA content axis in Figure 4.3-1. The number of cells for each point can be read from Figure 4.3-1. If the rate of increase of DNA content is constant throughout S, then the "ideal" DNA histogram would be two spikes flanking a truncated exponential. However, if the rate of DNA synthesis is not constant (which is usually the case), regions of relative high rates would have relatively few cells. The converse is also true. Thus, by analyzing the shape of the S phase in an "ideal" DNA histogram, we can obtain information on the relative average rate of DNA synthesis. The actual calculations for these rates can be found in Chapter 7.623. It should be emphasized that one can only obtain relative rates and not absolute rates from these calculations. There is absolutely no time information in a single DNA histogram. There is, however, a great deal of information in the distribution of cells in each of the cell cycle phases.

Figure 4.4-1: "Ideal" DNA histogram. The two relations: DNA vs cell age and number of cells vs cell age will give a "ideal" histogram when one solves them for number of cells vs DNA content. The first spike represents cells in G0 and G1 and the second spike represents cells in G2 and M. In between these two spikes are cells at various stages of DNA synthesis, S phase.

IDEAL HISTOGRAM



4.5 "Real" DNA histogram.

If we had a procedure to directly obtain an "ideal" histogram, this manuscript would be considerably shorter and a number of flow cytometetric computer experts would be out of jobs. The problem is when we actually try to measure the DNA content of cells, a number of uncertainties are introduced. For example, when cells are stained with a fluorescent probe such as propidium (PI), the actual ratio of number of PI molecules to number of nucleotides varies from cell to cell and from moment to moment. The DNA in cells is constantly twisting and expanding with Brownian movement. When the DNA expands, PI molecules fit in between the base pairs. When it contracts, PI molecules are forced out. Thus, at the precise moment of laser stimulation, there is going to be a small variation in the number of bound PI molecules and thus a variation in absolute fluorescence. There are a number of other areas where uncertainty can be introduced (e.g. unconstant sheath velocity, cell orientation and position, optics, and electronics). Each of these sources of error tend to make each point in the "ideal" DNA histogram a probability distribution. The precise nature of this probability distribution is a controversial issue and will be addressed more fully in Chapter 7.3. For now, let us assume the error is distributed normally (e.g. Gaussian distribution) and the DNA structure is identical in the various cell cycle phases. Let us also assume that the only source of uncertainty is the equilibrium of PI with DNA. The parameter that describes the spread of the probabiltiy distribution is called the standard deviation. We would first like to know how this standard deviation changes as function of DNA content. Figure 4.5-1A shows the highly idealized case of 2C amount of DNA combining with 10 molecules of PI. Let us say the standard deviation for this system is two. Now, let us double the amount of DNA. The total number of bound PI molecules would naturally be 20 and the standard deviation would be 4. Thus, by doubling the amount of DNA we double the standard deviation (see Figure 4.5-1B). Therefore, the relationship between DNA content (D) and standard deviation (SD) is given by

$$SD = CV * D \qquad (4.5\text{-}1)$$

Instead of using SD as an index of variation, it would make more sense to use an index that did not vary with DNA content. If one normalizes the standard deviation by dividing by the DNA content, the result will be a single number (CV) that will describe the system's variability. Since the standard deviation is frequently small in comparison to the DNA content, this number would be a fraction such as 0.05. If one multiplies this number by

100, the result is what is called the percent coefficient of variation (%CV) which is defined mathematically as

$$\%CV = SD * 100 / D \qquad (4.5\text{-}2)$$

If we now go back to the "ideal" DNA histogram and assume the above error distribution for each point, we obtain the "real" DNA histogram. Figure 4.5-2 shows the whole process of starting with number of cells vs cell age (Panel A) and DNA content vs cell age (Panel B) and ending up with the "ideal" DNA histogram (Panel C). When noise is introduced, we obtain the "real" DNA histogram. This figure is color coded in order to appreciate the fate of all the cell cycle phases. The fact that the S phase in the "ideal" histogram is compartmentalized into rectangles will be discussed more fully in Chapter 7.

The problem in analyzing DNA histograms is stripping away this noise and ending up with the "ideal" histogram or kernel. This process has the general term of deconvolution. Deconvolving histograms is necessary in order to evaluate the number of cells in each phase and especially to determine the relative rates of DNA synthesis throughout S. If one were to simply integrate the DNA histogram regions corresponding to G0-G1, S, and G2-M to find their percentages, the estimates would generally be extremely inaccurate. One could be off by as much as a factor of 2 or 3 in the estimation of S. The reason for this lies at the boundaries G0-G1/S and S/G2-M. These regions are essentially a "no man's land" for this type of calculation because there is no way of knowing whether you are counting S phase cells or not.

Before we introduce the methods for deconvolving DNA histograms, we will discuss the process by which an "ideal" histogram becomes a "real" histogram. As expected, this process is called convolution. The quantitation of the amount of information contained in a "real" histogram is the subject of the next chapter.

Figure 4.5-1A,B: Ideal relation between standard deviation and DNA content. A) Ten molecules of PI combine with 2C amount of DNA with a standard deviation of two. B) Twenty molecules of PI combine with 4C amount of DNA with a standard deviation of four. If one normalizes the standard deviation by dividing by the DNA content, the result is the coefficient of variation (CV) which is constant for this model.

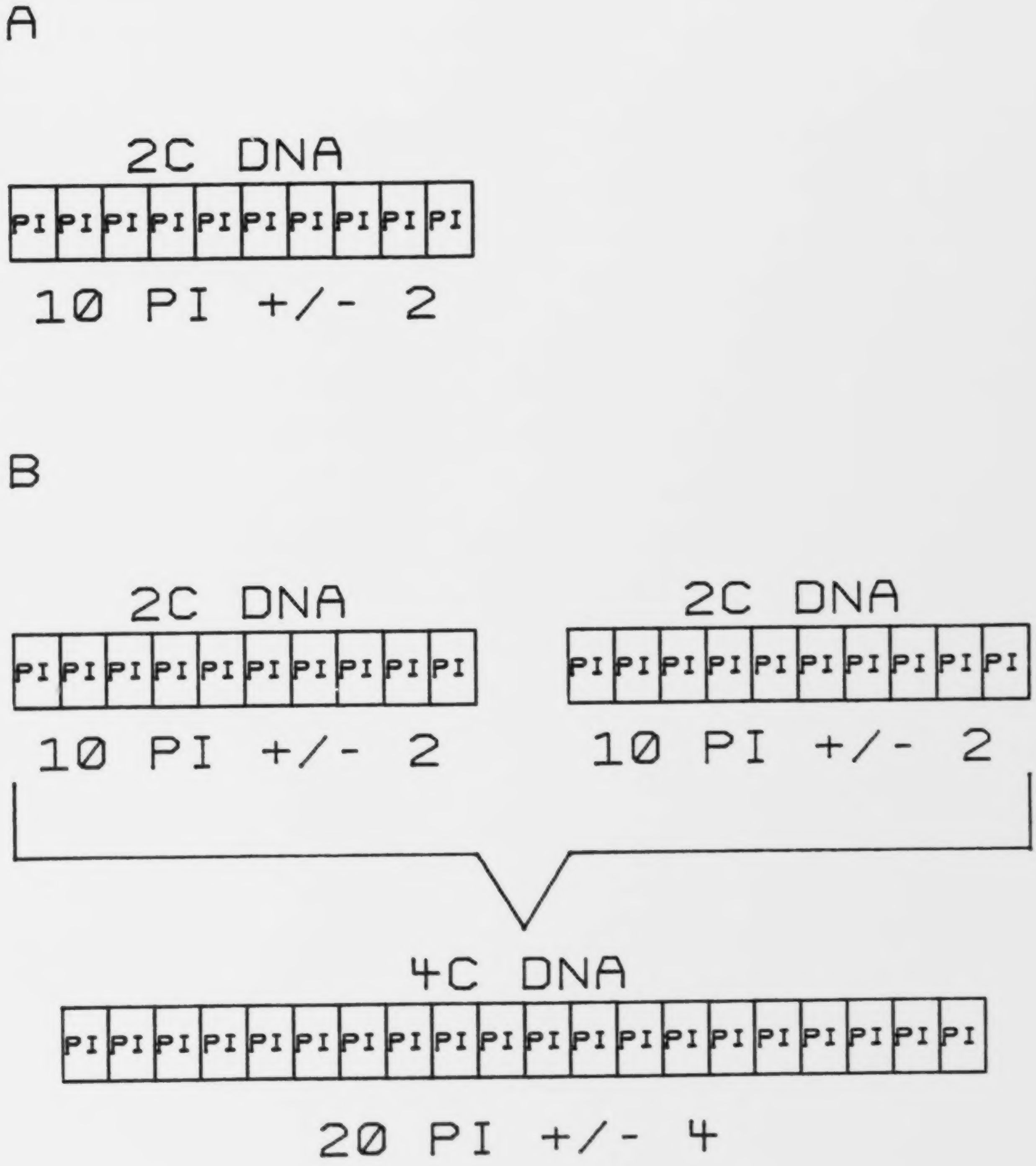
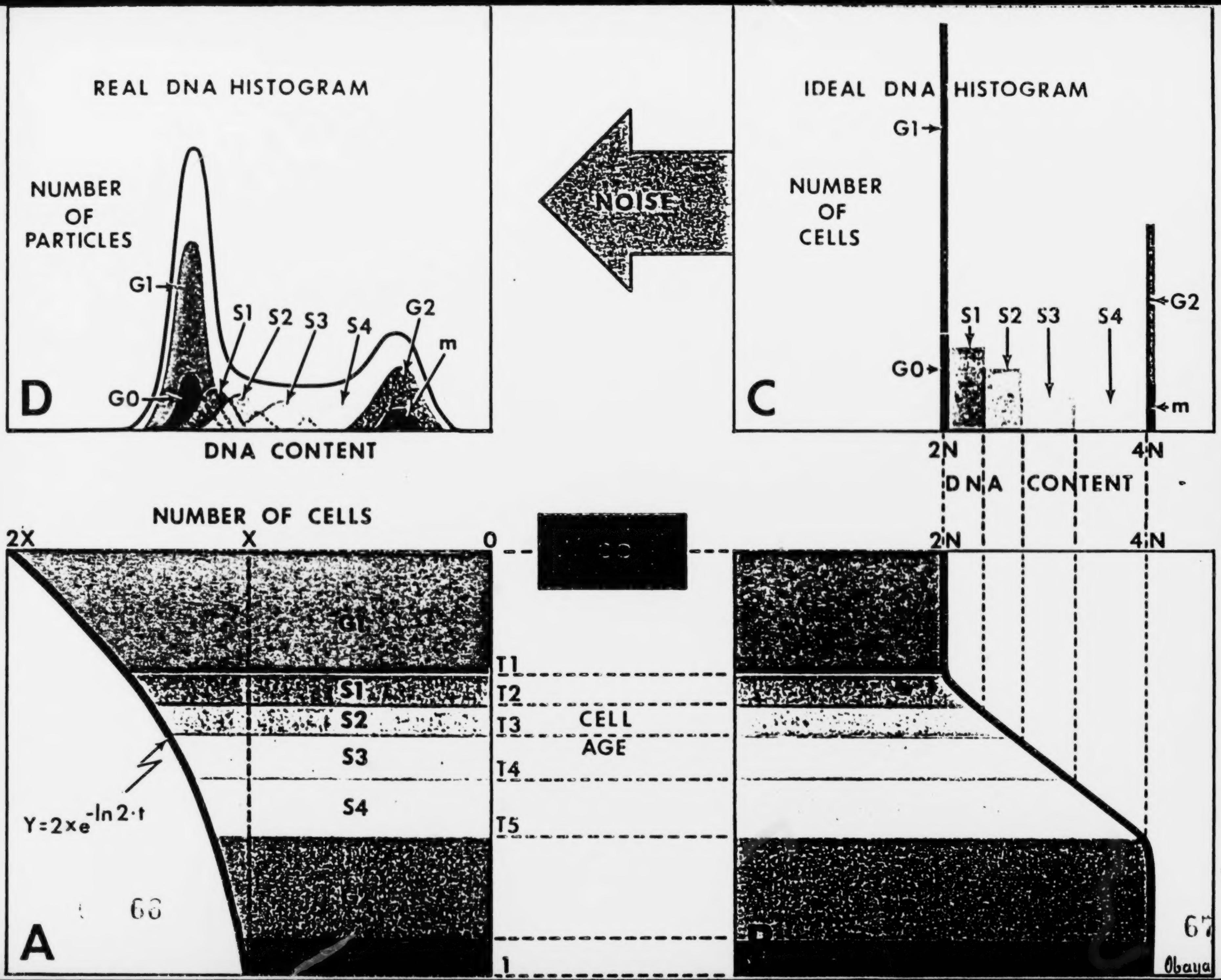

A
2C DNA
PI PI PI PI PI PI PI PI PI PI
10 PI +/- 2
B
2C DNA
PI PI PI PI PI PI PI PI PI PI
10 PI +/- 2
2C DNA
PI PI PI PI PI PI PI PI PI PI
10 PI +/- 2
4C DNA
PI PI PI PI PI PI PI PI PI PI PI PI PI PI PI PI PI PI PI PI
20 PI +/- 4

Figure 4.5-2: "Real" DNA histogram and its synthesis. A) Exponential relationship between number of cells and cell age. B) Relation between DNA content and cell age. C) "Ideal" DNA histogram. D) "Real" or observed histogram with uncertainty in the measurement of DNA content. This figure is color coded to aid in the appreciation of the fate of all the cell cycle phases.


REAL DNA HISTOGRAM
NUMBER OF PARTICLES
G1
G0
S1
S2
S3
S4
G2
m
D
DNA CONTENT
NOISE
IDEAL DNA HISTOGRAM
G1
NUMBER OF CELLS
S1
S2
S3
S4
G2
G0
m
C
2N
4N
DNA CONTENT
NUMBER OF CELLS
2X
X
0
G0
2N
4N
G1
S1
S2
S3
S4
T1
T2
T3
T4
T5
1
CELL AGE
Y=2xe^{-ln 2·t}
A
B
Obaya

5. Information theory.

5.1 Introduction.

The interpretation of a DNA histogram can be thought of as a transmission and receiving system as shown in Figure 5.1-1. We start with a large population of cells having precise DNA contents. The cells are then prepared for transmission by subjecting them to a probe such as propidium iodide (PI) which intercalates between the DNA base pairs stoichiometrically. A prepared cell is then selected at random, pushed into the core of a fast moving sheath and exposed to intense light. The PI fluoresces and the fluorescent photons travel from the cell to a detector where they are focused, filtered, converted to voltage pulses, and recorded on a multichannel analyzer. This process is repeated many thousands of times (eg. 10,000-100,000) and the final output is a "real" DNA histogram. The observed histogram is analyzed by a computer algorithm to yield the most probable set of original messages for a given kernel model. In any transmission process, one inevitably will lose information. The science of mathematics that attempts to quantitate this loss is embodied in information theory or more pessimistically, entropy theory.

Instead of beginning with the formal definition of information theory, let us begin with a simple example and slowly add complexity to it. Suppose we were presented with four sealed boxes of uniform size (15). Inside one box is a ball. We have no a priori knowledge of where that ball is. It has an equal probability of being in any one of the four boxes. Therefore, the probability of it being in any one of the boxes is 1/4 or 0.25. Now, suppose we were asked how many yes/no type questions would it take, on the average, to find the ball. After a little consideration one would naturally say, two. A diagram showing how the questions might be constructed is shown in Figure 5.1-2. In this particular case it will always take two questions to find the ball. The amount of uncertainty or entropy (H) associated with this system is actually the minimum average number of questions it takes to find the ball. It can be shown (16) that this uncertainty is given by

$$H(X) = \sum_{i=1}^{n} P(X=i) * \mathrm{LOG}_{2}(1/P(X=i)) \qquad (5.1\text{-}1)$$

where X is the set of n boxes; P(X=i) is probability of the ball being in box i; and H(X) is the average uncertainty of the system. If the log is to the base 2, then H(X) is in units of BITS. If the log is to the base e, then H(X) is in units of NITS. One can easily convert from NITS to BITS by dividing the NITS by LN(2). Let us see if this formula works for our simple case.

$$H(X) = \sum_{i=1}^{4} \frac{1}{4} * \mathrm{LOG}_2(4) = 2 \text{ BITS} \qquad (5.1\text{-}2)$$

Since a BIT is equivalent to the answer to a yes/no question, this formula gives us the correct answer. Equation 5.1-2 will always give the minimum average entropy for a particular system. Let us now complicate our example slightly. Instead of the four boxes being uniform in size, we have three equal size boxes and one large box equal in size to the sum of the other three (see top of Figure 5.1-3). If the chance of finding the ball in a particular box is proportional to its size, then the probabilities for this system would be 1/6, 1/6, 1/6, and 1/2 (see Figure 5.1-3). The question is, on the average, how many yes/no questions would it take to find the ball now. If we apply equation 5.1-1 to this problem we obtain

$$H(X) = \frac{1}{2} * \mathrm{LOG}_2(6) + \frac{1}{2} * \mathrm{LOG}_2(2) = 1.79 \text{ BITS} \qquad (5.1\text{-}3)$$

Figure 5.1-3 shows how the questions might be optimally configured. We can actually calculate the average number of questions associated with this system as

$$\text{Av num of questions} = .5(1) + .5(\,(1/3)*2 + (2/3)*3) = 1.83 \qquad (5.1\text{-}4)$$

The actual average of yes/no questions (1.83) is slightly higher than the calculated uncertainty (1.79). This will generally be the case because of the physical inability to optimally configure a question scheme. An optimal yes/no decision scheme is where the answer has an equal probability of being yes or no at every level. You can see in Figure 5.1-3 that the question: is the ball in box 1 (1?) had a 1/3 chance of being yes and a 2/3 chance of being no; thus deviating slightly from optimality.

Information is the reduction of uncertainty after some observation. In our last example, suppose instead of asking questions we were allowed to see through the boxes and the ball happened to be in box 2. The uncertainty as to where the ball is after this observation would, of course, be zero. Therefore, the information gained would be

$$I(X/Y) = H(X) - H(X/Y) = 1.79 - 0 = 1.79 \text{ BITS} \quad (5.1\text{-}5)$$

where $I(X/Y)$ is the information gained about X after having observed Y; $H(X)$ is the original uncertainty of X; and $H(X/Y)$ is the uncertainty of X after having observed Y. Now, suppose that although we thought we observed the ball in box 2, we really were not sure whether it was in box 2 of box 3 (it was a long night). Therefore, after observing Y, the probabilities for boxes 1 through 4 are 0, .5, .5, 0 respectively. Notice that now it would take only 1 yes/no question to find the ball. This is expressed mathematically as

$$I(X/Y) = H(X) - H(X/Y) = 1.79 - \log_2(2) = .79 \text{ BITS} \quad (5.1\text{-}6)$$

The concept is conceptually easy. If our original system takes on the average H(X) questions to find the ball and after some kind of observation (Y), it only takes H(X/Y) questions to find the ball, then the amount of information gained about the system after the observation is H(X) - H(X/Y).

We will now add a little more complexity to the model. Let us assume that instead of having a single ball (single message), we have a number in each box (multiple messages). The only a priori knowledge we have is that the total of all the numbers is Z or in this case, 5. The uncertainty for this system is going to be a function of the total number of possible combinations of these numbers summing to Z. Table 5.1-1 gives all the combinations for this example. We can see from this table that the total number of possibilities is 56. We could have obtained this number directly by using the combination formula

$$C(X;Z) = \frac{(Z+X-1)!}{Z!\ (X-1)!} \qquad (5.1\text{-}7)$$

where C(X;Z) is the number of combinations; Z is the number of events; and X is the number of boxes. Let us check it out for our example. When Z=5 and X=4 we have

$$C(X;Z) = \frac{8*7*6*5*4*3*2}{(5*4*3*2)*(3*2)} = \frac{8*7*6}{3*2} = 56.$$

This formula will work for all combinations of Z and X.

Table 5.1-1: All combinations of four numbers summing to five.

0005
0014
0023
0032
0041
0050

0104
0113
0122
0131
0140

0203
0212
0221
0230

0302
0311
0320

0401
0410

0500

1004
1013
1022
1031
1040

1103
1112
1121
1130

1202
1211
1220

1301
1310

1400

2003
2012
2021
2030

2102
2111
2120

2201
2210

2300

3002
3011
3020

3101
3110

3200

4001
4010

4100

5000

TOTAL = 56

The problem with equation 5.1-7 is that for large Z's and X's, the computation becomes exceedingly tedious. Fortunately, there exists a very good estimation of the factorial, the Sterling approximation (17).

$$N! \sim (2*Pi*N)^{1/2} * N^{N} * e^{-N} \quad \text{or}$$

$$LN(N!) \sim -N + N*LN(N) + 1/2*LN(2*Pi*N) \qquad (5.1\text{-}8)$$

Therefore, C(X;Z) can be approximated by

$$C(X;Z)=EXP\left[(Z+X-1)LN(Z+X-1)-ZLNZ-(X-1)LN(X-1)+\frac{1}{2}LN\frac{Z+X-1}{2PiZ*(X-1)}\right]$$

$$(5.1\text{-}9)$$

If we plug our values of Z and X into this formula we get 57.93 which is close to the actual value of 56. Actually as Z and X become large, the accuracy of this approximation becomes much better. Let us assume that each of the 56 combinations have an equal chance of occurring, then the average entropy for this system is

$$H(X) = \sum_{i=1}^{56} P(X=i) * LOG_2\left(\frac{1}{P(X=i)}\right) = \log_2 C(X;Z) = 5.81 \text{ BITS}$$

that is, on the average, it would take a minimum of 5.81 yes/no questions to determine the actual numbers in the boxes. If we were allowed to see through the boxes, we might read a number such as 1121. The information gain would obviously be

$$I(X/Y) = H(X) - H(X/Y) = 5.81 - 0 = 5.81 \text{ BITS}$$

We might interpret this as meaning that by observing the contents of boxes we were able to save ourselves, on the average, 5.81 yes/no questions.

We assumed in this example that all 56 combinations had an equal weight. What if this were not true? Suppose that for some reason, numbers with at least two zeros were twice as likely to occur than numbers with less than 2 zeros. If we give the 28 numbers with two zeros or more a weighting factor of 2 and the others a weighting factor of 1, then we can recompute the probabilities for each number as

$$Pi = \frac{Wi}{\sum_{i=1}^{n} Wi}$$

Thus, the two or more zero numbers have a probability of

$$Pi = \frac{2}{84}$$

and the others have a probability of

$$Pi = \frac{1}{84}$$

The recomputed entropy for this system would be

$$H(X) = \sum Pi * \log_2 \left(\frac{1}{Pi} \right)$$

$$= \frac{28*2}{84} * \mathrm{LOG}_2 \left(\frac{84}{2} \right) + \frac{28}{84} * \log_2 (84) = 5.72 \text{ BITS}$$

As expected because some combinations were more likely than others, it took, on the average, less yes/no questions to find any number. In fact it can be proven that the entropy of a system is at its maximum when the probability of all its elements are equal (16). Operationally, the way we would organize our questions would be to put all the combinations down in descending probability. We would then separate them into two unequally sized groups, each with about a .5 chance of containing the true combination. We then ask the yes/no question as to which group its in. We continue this procedure until we find the true combination. On the average, the best we could hope for is 5.72 questions.

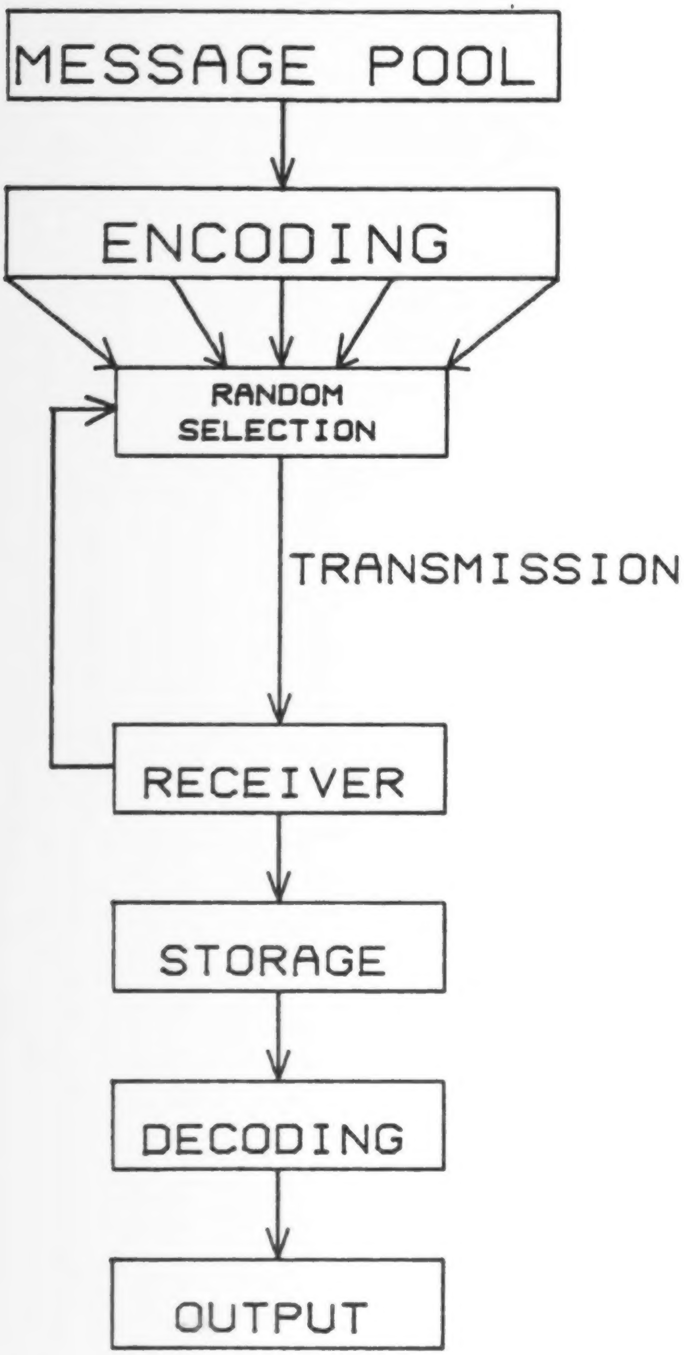
There are some conditions that will greatly reduce the number of calculations necessary to find the entropy of a system. Suppose that we were given the information that our number did not contain any zeros. By looking at Table 5.1-1 we see that only four numbers do not contain any zeros; therefore, that piece of information was worth

$$I(X/Y) = 5.81 - \sum_{i=1}^{4} \frac{1}{4} * \log_2(4) = 5.81 - 2 = 3.81 \text{ BITS}$$

The major point to notice is that if a combination has a zero weight or probability, it is not involved in the calculation.

The calculation of structural information content of a signal such as a DNA histogram involves the application of the preceding principles to both the signal and to random statistical noise.

Figure 5.1-1: DNA histogram transmission process.


MESSAGE POOL
ENCODING
RANDOM SELECTION
TRANSMISSION
RECEIVER
STORAGE
DECODING
OUTPUT

BIOLOGIC INTERPRETATION

A population of cells each having a particular DNA content.

The cells are stained such that a fluorescent probe (eg. PI) stoichiometrically intercalates into the DNA.

The cells are pushed into the core of a fast moving sheath and exposed to light excitation.

A small percentage of emitted fluorescent photons travel from the cell to a detector where it is converted to a voltage pulse. This process is repeated many thousands of times.

The encoded message is stored in a multichannel analyzer.

The collection of encoded messages, the DNA histogram, is analyzed by some computer algorithm.

The final output is the most probable set of original messages.

Figure 5.1-2: Optimal yes/no question structure for finding a ball in four equally probable boxes. The symbol / is representative of the word "or".

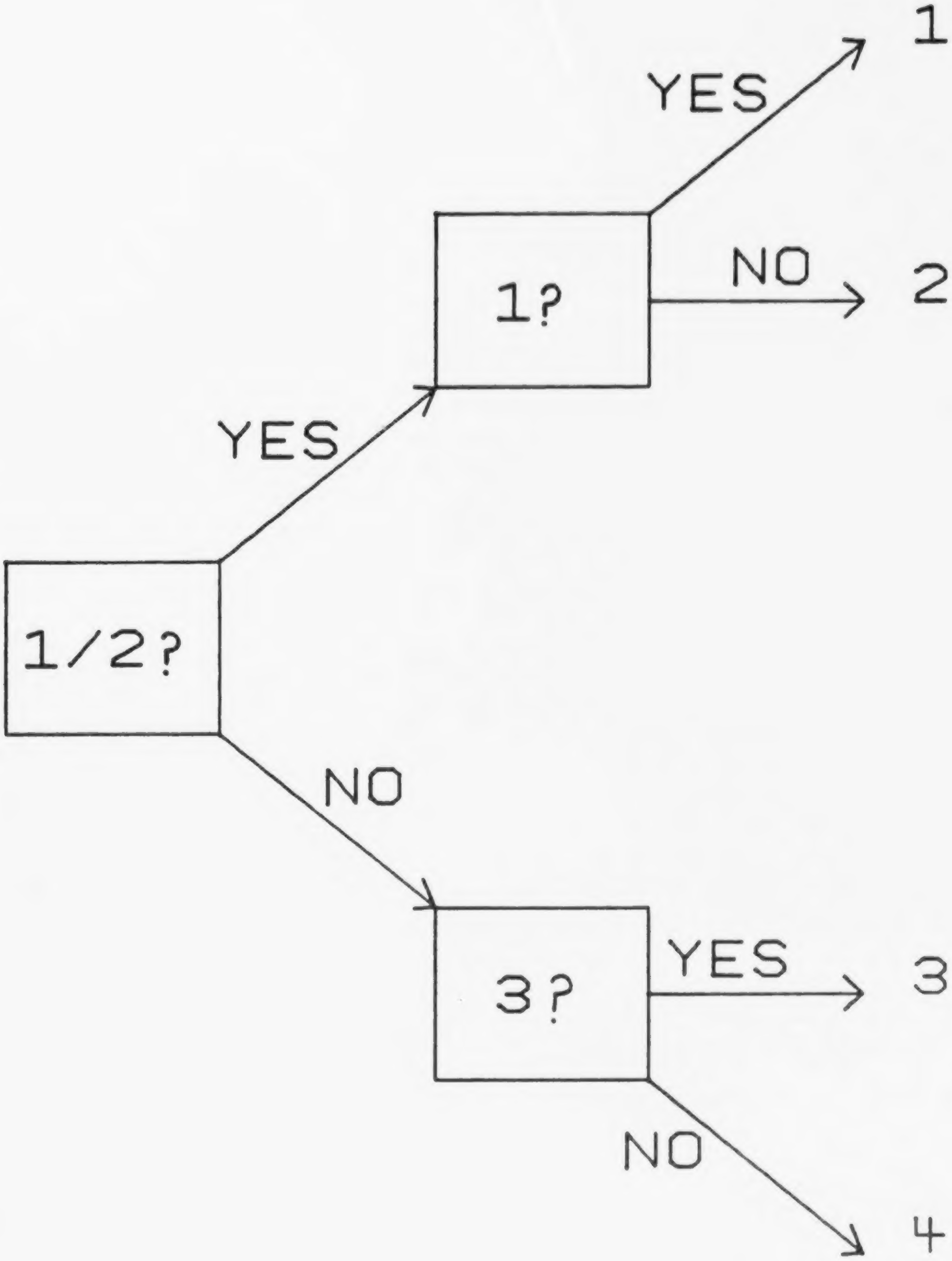
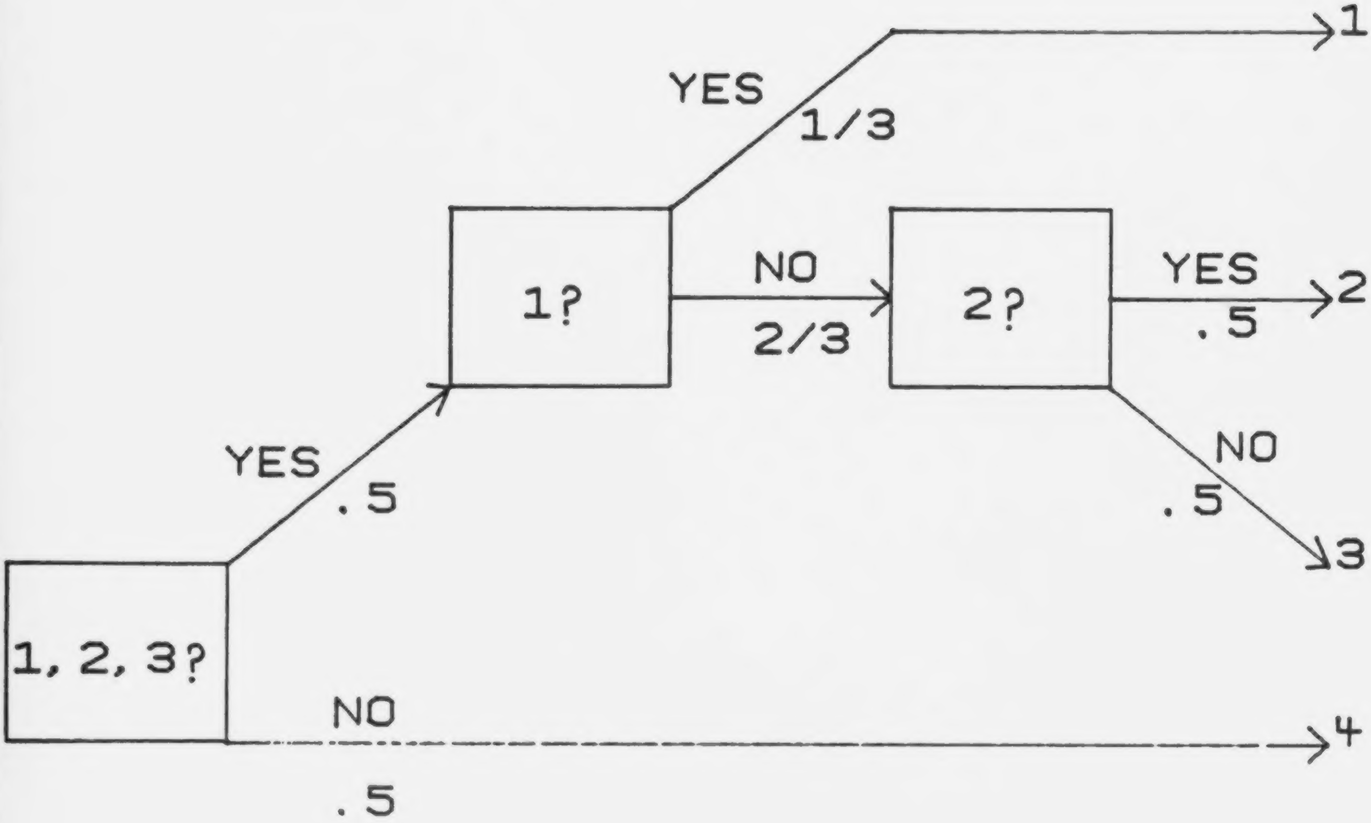

1/2?
YES
NO
1?
YES
NO
3?
YES
NO
1
2
3
4

Figure 5.1-3: Optimal yes/no question structure for finding a ball in unequally probable boxes: 1/6, 1/6, 1/6, and 1/2.


1, 2, 3?
YES
.5
NO
.5
1?
YES
1/3
NO
2/3
2?
YES
.5
NO
.5
1
2
3
4

5.2 Flow cytometric structural information content.

If we had an ideal 128 channel cytometer and a signal with 50,000 totally independent messages with no noise, the entropy associated with such a system would be

$$H(X) = LOG_{2}(C(128;50,000)) = 1269 \text{ BITS}$$

Another way of looking at this is if we asked one yes/no question a second, it would take on the average, 21.15 minutes to find the correct sequence of numbers that made up the true histogram. Let us suppose for a moment that the cytometer is constructed such that each channel has an equal probability of receiving a message. Let us further assume that this cytometer has 128 channels and the total accumulation in these channels is 10,000. What would we observe on the flow cytometer if there were absolutely no biologic structure, that is, a total random signal? Since there is an equal probability of each channel containing counts, we would expect that the 10,000 counts would be somewhat equally dispersed over the 128 channels. We also would expect that, on the average, a given channel would contain 10,000/128 or 78.12 counts per channel. There would, however, be a variation about this mean value. The variation obeys Poisson statistics for counts greater than ten (17) and the standard deviation for the distribution is simply the square root of the mean value. A very good approximation to the Poisson for large counts is the Gaussian distribution. The two free parameters that describe the Gaussian distribution are its mean and standard deviation. Thus, for our simple example, we would expect counts distributed about a mean value of 78.12 with a standard deviation of SQR(78.12) or 8.84. The continuous Gaussian probability distribution is given by

$$P(X;M,SD) = \frac{1}{SQR(2*PI)*SD(N)} * e^{-.5*((X-M)/SD(N))^{2}} \, dX \qquad (5.2\text{-}1)$$

where P is the probability of finding X within the interval dx with a parent distribution described by a standard deviation of SD(N) and a mean of M. We can ask the question what is the uncertainty associated with such a set of probabilities? We can compute the uncertainty directly by applying equation 5.1-1, that is

$$H(X) = \sum_{i=1}^{n} P(Xi;M,SD(N)) * LOG_{2} \left(\frac{1}{P(Xi;M,SD(N))} \right) \qquad (5.2\text{-}2)$$

This computation has two serious limitations. First, since the Gaussian is defined for all real numbers (including negative numbers), the limits of the summation should be from minus infinity to plus infinity. The second limitation is that practically this calculation would take an enormous amount of time. Therefore, it would be advantageous to put equation 5.2-2 in closed form. This is done by evaluating the integral

$$H(X) = \int P(X;M,SD(N)) * \mathrm{LOG}_2 \left(\frac{1}{P(X;M,SD(N)} \right) dX$$

where P(X;M,SD(N)) is defined in equation 5.2-1. The derivation is slightly complex, but the result is delightfully simple.

$$H(X;SD)) = .5 * \mathrm{LOG}_2 (2*PI*e*SD^2) \quad (5.2\text{-}3)$$

Now, let us get back to our example. The entropy associated with a Gaussian with a standard deviation of 8.84 is easily evaluated by equation 5.2-3 as 5.19 BITS. Therefore, on the average, it would take us 5.19 BITS to find the correct channel accumulation if its probability distribution could be approximated by a Gaussian with a standard deviation of 8.84. If we followed this particular search pattern for each channel in our 128 channel cytometer, the total amount of BITS needed to find the correct sequence for the random signal would be 128*5.19 or 664.32 BITS. Figure 5.2-1 shows the random signal for this particular configuration along with its Gaussian probability distribution. The results of the mean, standard deviation, and entropy calculations are located in the bottom right hand corner.

What happens when we have a true signal? Let us assume that we have a sine wave for a signal with a period that extends over 128 channels. Let us have the area underneath this sine wave equal to 50,000 counts. Each point along the sine wave function is actually a probability distribution with a standard deviation equal to the square root of the noiseless value. In the same manner as with the random signal, we can compute the probability distribution that describes the entire sine wave signal.

$$M(S) = \sum_{i=1}^{128} Yi \;/128$$

$$SD(S) = SQR \left[\sum_{i=1}^{128} \left(Yi - M(S) \right)^{2} /127 \right]$$

At the same time, we can compute the probability distribution for a random signal with a mean equal to $M(S)$ and a standard deviation equal to $SQR(M(S))$. Figure 5.2-2 shows both these signals with their corresponding probability distributions. We can compute the entropy associated with the sine wave signal by applying equation 5.2-3. The standard deviation for this particular case was 212.14 and the calculated uncertainty was 9.77 BITS. The average number of yes/no questions needed to determine a particular accumulation would be 9.77 BITS. However, it takes 6.34 BITS to find the accumulation for a totally random signal. Therefore, the net number of BITS needed to identify the true structural information of the signal for one channel is 9.77-6.34 or 3.43 BITS. The total amount of structural information would be 128*3.43 BITS or around 439 BITS. This process can be described mathematically as

$$I(X/S) = .5C*\mathrm{LOG}_2(2PIe*SD(S)^2) - .5C*\mathrm{LOG}_2(2PIe*SD(N)^2)$$

$$= .5*C*\mathrm{LOG}_2\left(\frac{SD(S)^2}{SD(N)^2}\right)$$

$$= \frac{1}{4} * C * \mathrm{LOG}_2\left(\frac{SD(S)}{SD(N)}\right) \qquad (5.2\text{-}4)$$

This formula states that the average information content is a function of the ratio of the signal standard deviation to the corresponding noise standard deviation. One can gain a better appreciation of this formula by decreasing the amplitude of the sine wave in increments and calculating its information content. Figure 5.2-3 shows the information content of the sine wave signal with an amplitude of 75 to be only 186.6 BITS. When one decreases the amplitude to 30, there appears very little difference between the random signal and the sine wave; however, this method has enough sensitivity to compute an information content of 70.27 BITS (see Figure 5.2-4). Figure 5.2-5 summarizes the relationship between information content and sine wave amplitude. The positive relationship between sine wave amplitude and information content was expected, but the sensitivity of this assay system for small amplitudes was startling. Significant structural information was found for a sine wave with an amplitude of only 15 where the random noise standard deviation was 19.7.

In the beginning of this chapter we made the assumption that the probability of an event registering in any channel was equal. Unfortunately, this is not the case for most flow cytometers today. If one doubles the channel number, the standard deviation associated with that channel number also doubles. Stated mathematically, the probability of an event registering in its proper channel is

$$P(i) = \frac{1}{\mathrm{SQR}(2*\mathrm{PI}) * \mathrm{CV} * i} \tag{5.2-5}$$

where CV is the percent coefficient of variation and i is the channel number. Thus, if we were to observe a totally random set of messages with such a flow cytometer, it would be a decreasing hyperbolic function which would have an obvious deleterious effect on our calculations. There are at least two methods to solve this problem. A log amplifier can be inserted between the detector and multichannel analyzer. The channel probabilities (line spread) would be constant with this configuration. There are a number of other advantages with the log amplifier. A greater range of DNA ploidy can be observed. Cell types such as megakaryocytes can have a range of ploidy from 2C up to 64C. With a linear amplifier, it is next to impossible to quantitate all the peaks. The other advantage is that there is theoretically as much as twice as much information content in identical histograms when obtained with a log amplifier. This will be discussed later in chapter 7.3. The other alternative is to use computer software to transform the histogram. This method requires a rather complex mapping procedure and subsequent smoothing. The end result is a histogram with a constant line spread but with an unpleasant choppy appearance (see Figure 5.2-6). It is unlikely that the transformation has much of a deleterious effect on the final calculations since the fluctuations are being averaged over the entire 128 channels. The mapping equation used was designed to spare channel loss due to histogram compression in the log domain. Let X be the original intensity axis and V be the transformed axis, the equation that describes the map is

$$Vi = 1 + \frac{(C-1)*(LN(i) - LN(P/2))}{LN(C) - LN(P/2)} \quad \text{for } i=P/2 \text{ to } 128 \tag{5.2-6}$$

where i is the X axis channel number, P is the first peak's location and C is the number of channels. This map will result in an incredibly choppy histogram unless some modest smoothing is performed. The algorithm used here was

if $Vi+1 - Vi = 1$ then no smoothing.
if $Vi+1 - Vi > 1$ then distribute equally the counts destined for Vi over the interval $Vi \leq V \leq V(i+1) - 1$.
if $Vi+1 - Vi < 1$ then distribute the counts destined for Vi equally between Vi and Vi-1.

This method was chosen because of its simplicity and its mild smoothing characteristics (no count loss).

The technique of determining information content has a number of advantages and applications. The advantages over other approaches is that it is simple, intuitive, sensitive, and fast. The major disadvantage or limitation is the assumption that the signal can be approximated with a Gaussian probability function. The simplicity of equation 5.2-4 argues in its behalf, but a more accurate description of the overall probability distribution could probably be obtained by breaking the Y axis into a number of small intervals, forming a frequency histogram, and calculating the entropy using equation 5.2-2. We choose the simpler of these two alternatives, namely, the Gaussian approximation.

One possible application of this technique is in evaluation of new flow cytometric designs. Minimizing CV is not of paramount importance. What is important is to maximize the structural information content of some known standard such as beads. The structural information content is a function of not only CV but also the signal to background noise ratio. As the CV or background decreases, the SD(S) will tend to increase, thus raising the calculated information content. There is an almost fanatical attempt among manufacturers and investigators to reduce the CV in flow cytometers. The actual percent difference this makes on the information content is quite modest. Figure 5.2-7 shows a set of histograms with identical characteristics except for %CV which ranges from 1 to 7. An information analysis such as the one presented in Figure 5.2-6 was performed on each histogram. The information content vs %CV is summarized in Figure 5.2-8. As expected, information content increases with decreasing CV; however, a 5 fold diminution of %CV (5 to 1) resulted in only a 21% increase in information content.

One can use the principles described in this chapter to determine the information gain by using a log amplifier in the flow cytometer (Chapter 7.3). For normal operating conditions, the increase is approximately 150%, but for high accumulations it asymptotically approaches 200%. This difference is attributed to the inefficiency of linear cytometers. When the signal is of low intensity there are too few channels to accumulate the information and when the signal is of high intensity there are too many channels. The log cytometer, however, uses its channels equally since the line spread function is constant.

This chapter answers the question of how much structural information is actually contained in a DNA histogram. One can also use information theory to compute the amount of information contained in each channel of a flow cytometer. This quantity is called information density and is the subject of the next chapter.

Figure 5.2-1: Random signal with its Gaussian probability function for a 128 channel, 50,000 accumulation.

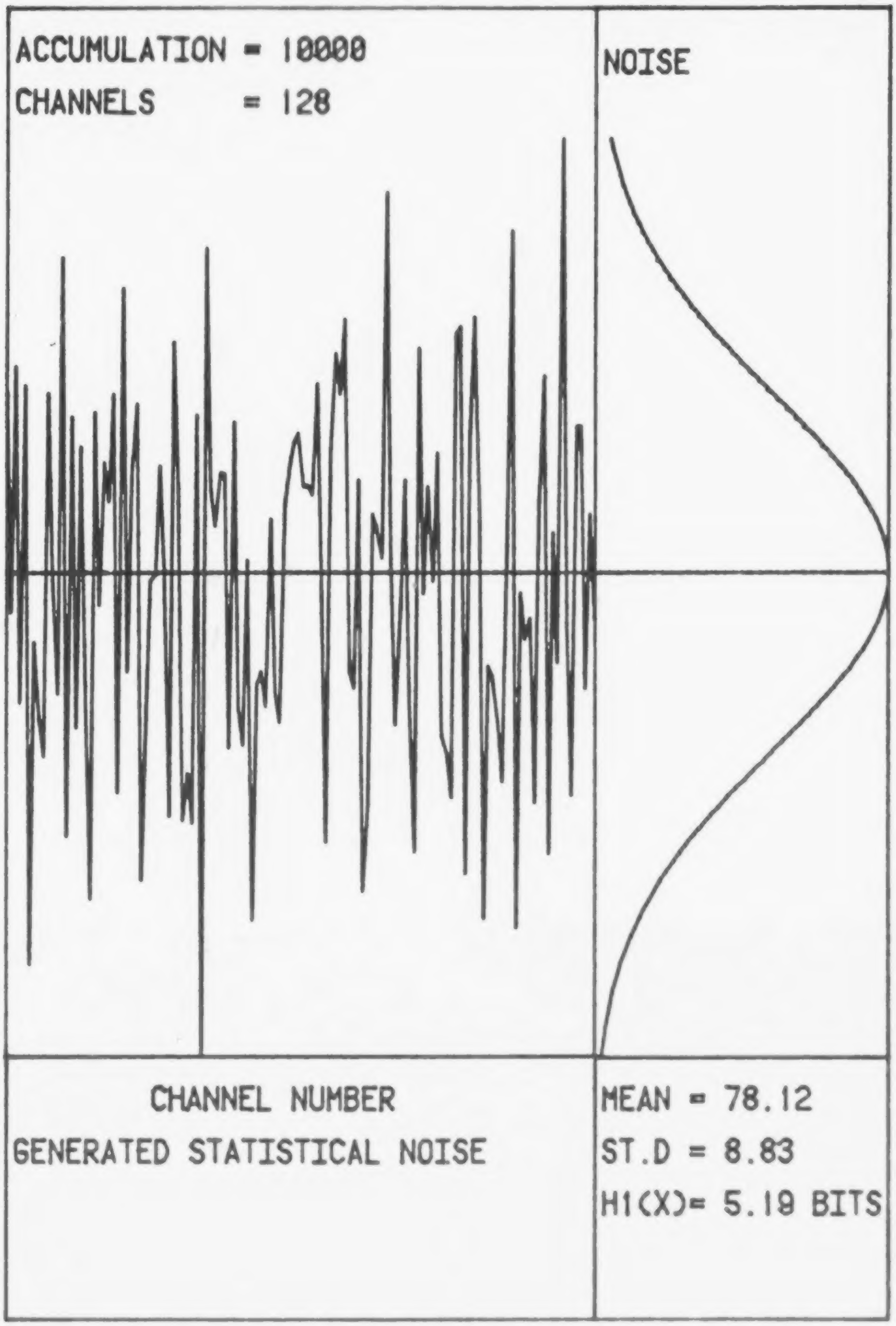

ACCUMULATION = 10000
CHANNELS = 128
NOISE
CHANNEL NUMBER
GENERATED STATISTICAL NOISE
MEAN = 78.12
ST.D = 8.83
H1(X)= 5.19 BITS

Figure 5.2-2: Information analysis of a sine wave with an amplitude of 300. Accumulation = 50,000 and number of channels = 128.

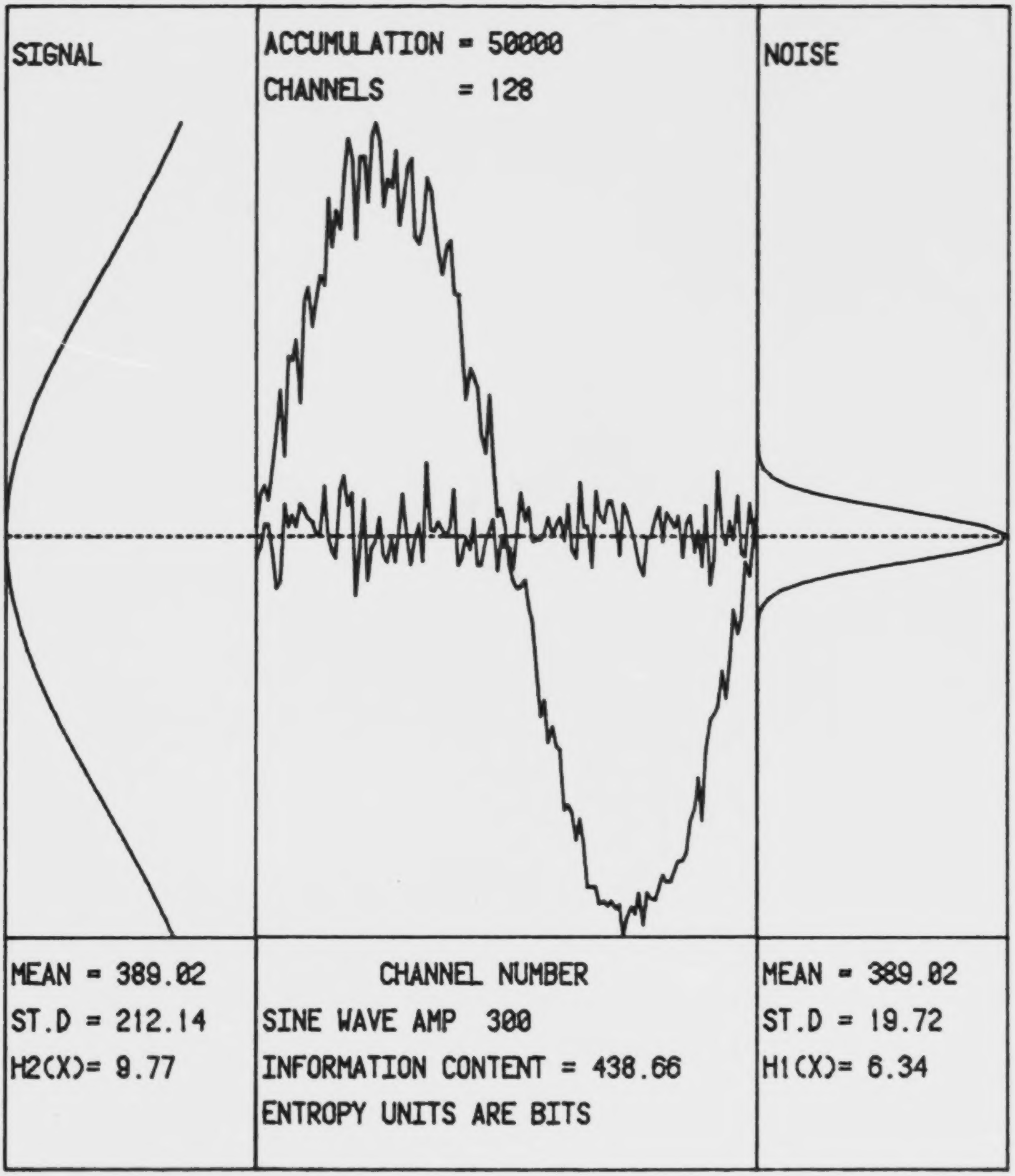

SIGNAL
ACCUMULATION = 50000
CHANNELS = 128
NOISE
MEAN = 389.02
ST.D = 212.14
H2(X)= 9.77
CHANNEL NUMBER
SINE WAVE AMP 300
INFORMATION CONTENT = 438.66
ENTROPY UNITS ARE BITS
MEAN = 389.02
ST.D = 19.72
H1(X)= 6.34

Figure 5.2-3: Information analysis of a sine wave with an amplitude of 75.

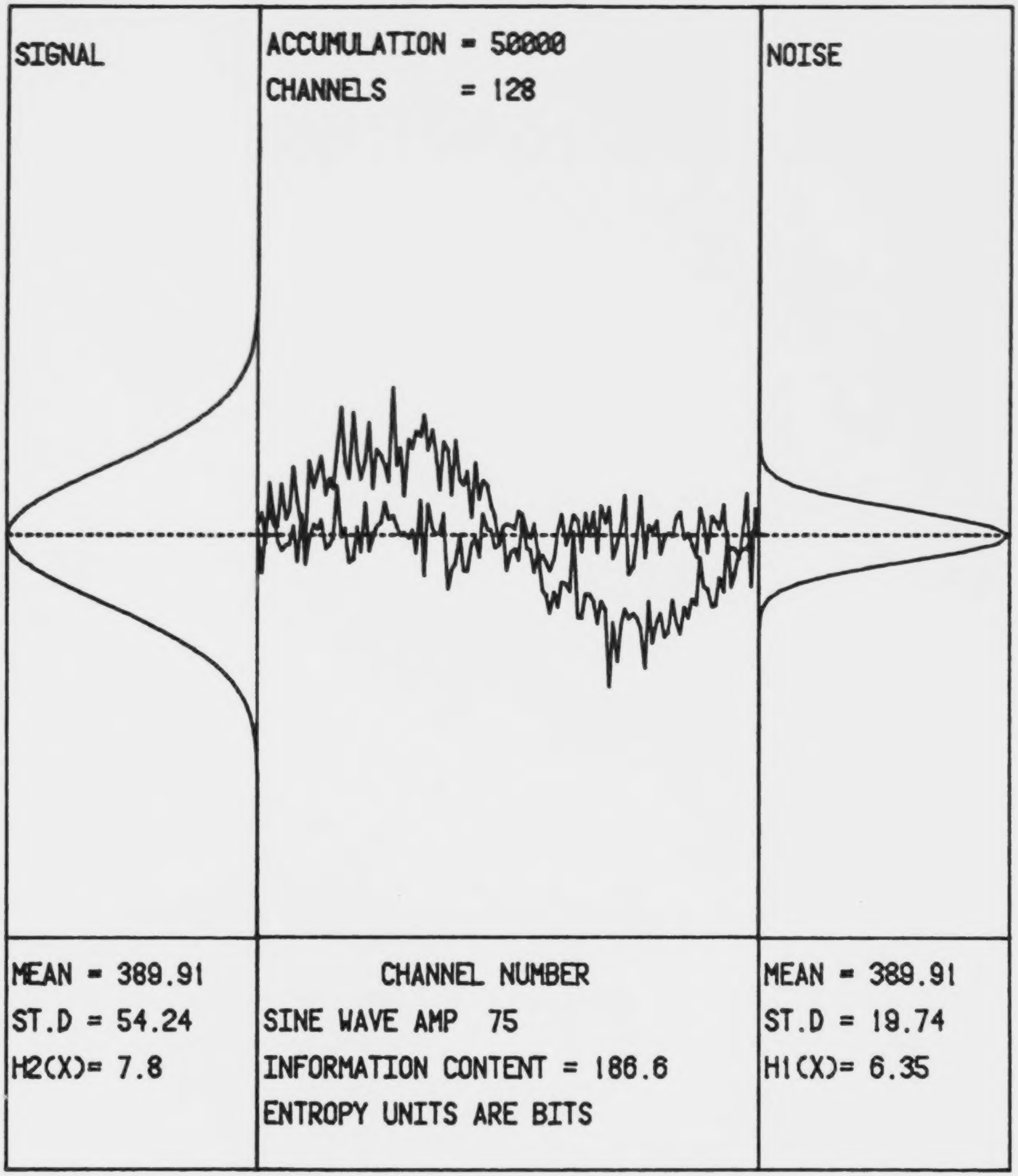

SIGNAL
ACCUMULATION = 50000
CHANNELS = 128
NOISE
MEAN = 389.91
ST.D = 54.24
H2(X)= 7.8
CHANNEL NUMBER
SINE WAVE AMP 75
INFORMATION CONTENT = 186.6
ENTROPY UNITS ARE BITS
MEAN = 389.91
ST.D = 19.74
H1(X)= 6.35

Figure 5.2-4: Information analysis of a sine wave with an amplitude of 30.

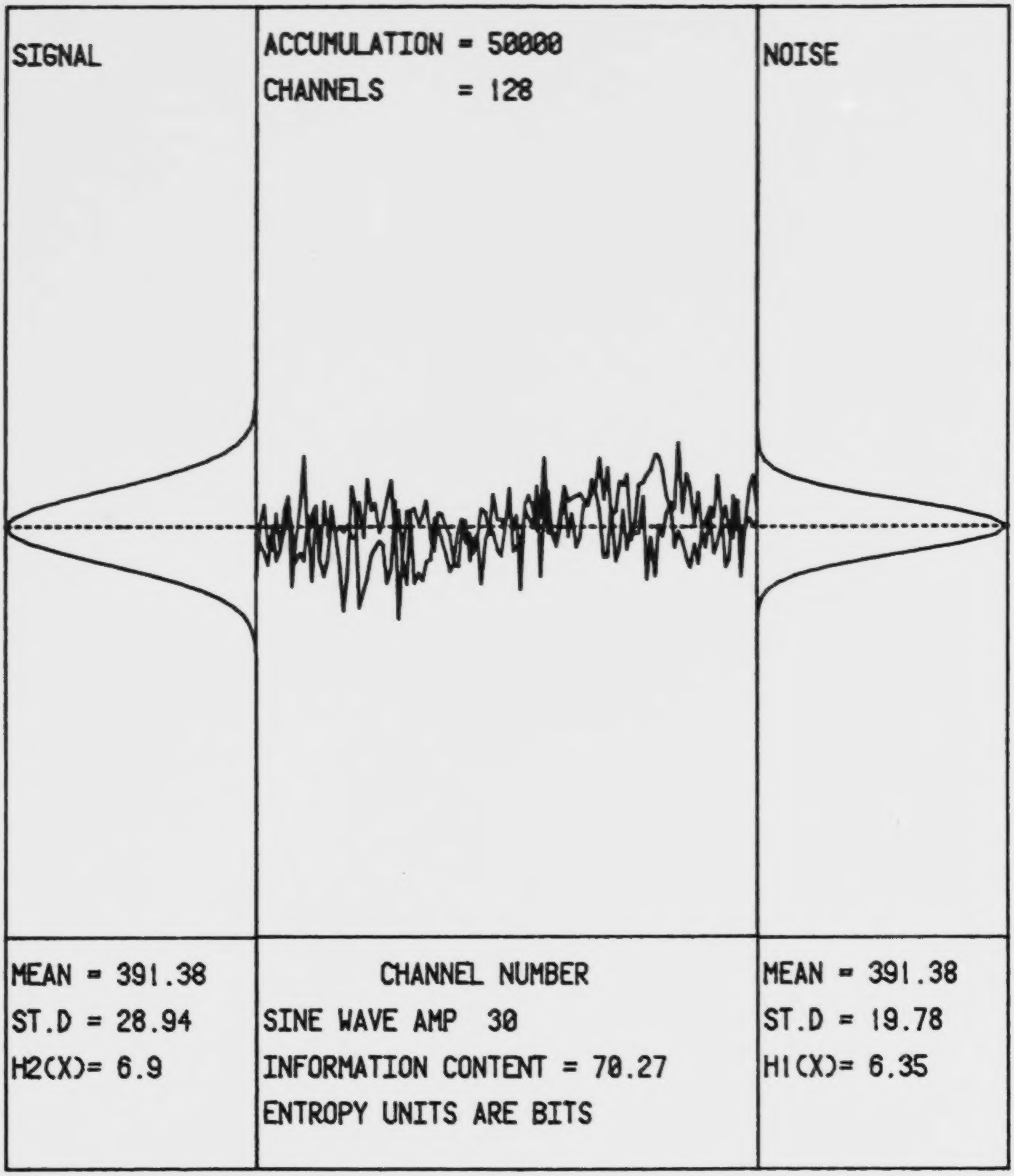

SIGNAL
ACCUMULATION = 50000
CHANNELS = 128
NOISE
MEAN = 391.38
ST.D = 28.94
H2(X)= 6.9
CHANNEL NUMBER
SINE WAVE AMP 30
INFORMATION CONTENT = 70.27
ENTROPY UNITS ARE BITS
MEAN = 391.38
ST.D = 19.78
H1(X)= 6.35

Figure 5.2-5: Summary of information content vs sine wave amplitude.

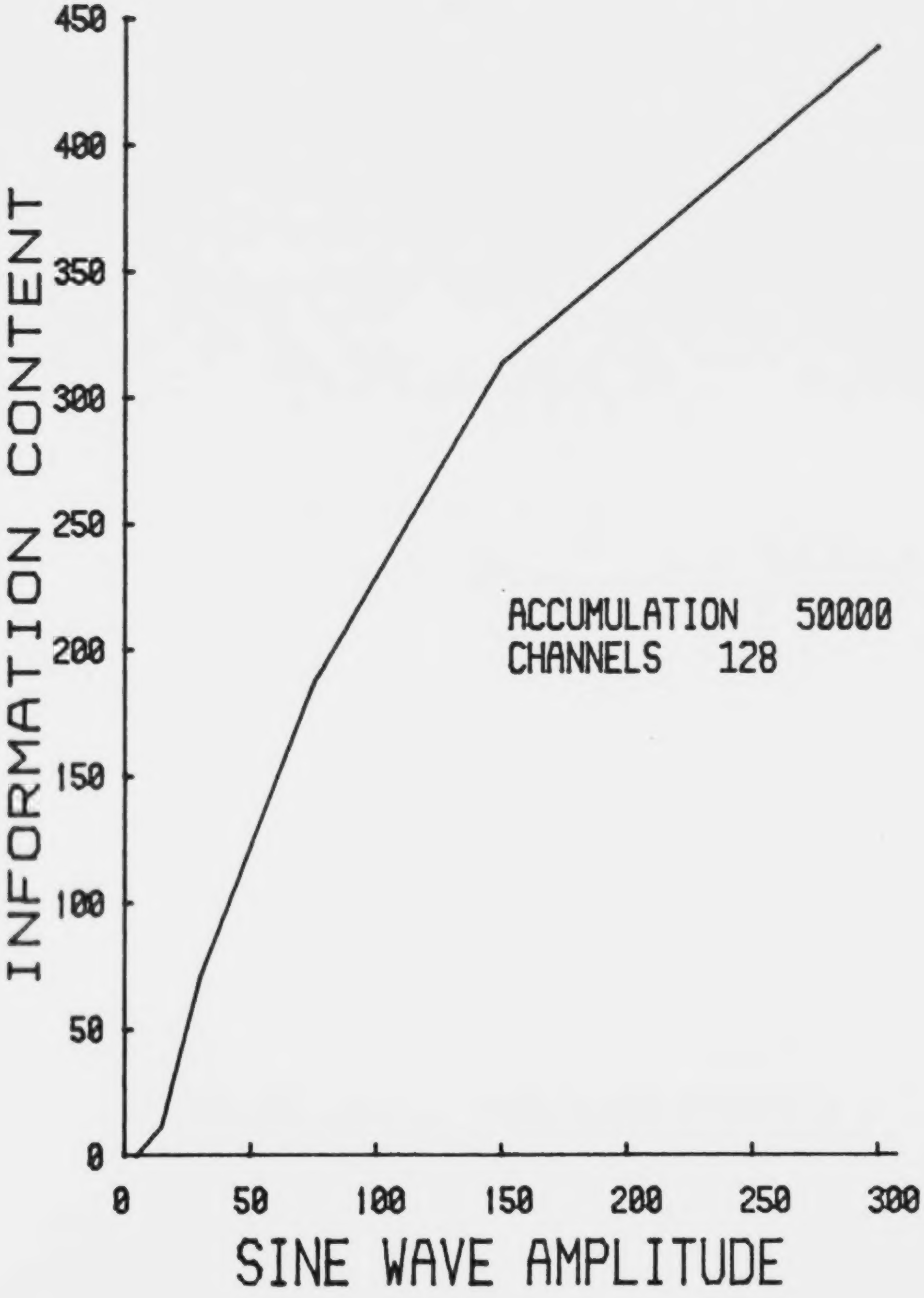

450
400
350
300
250
200
150
100
50
0
INFORMATION CONTENT
0
50
100
150
200
250
300
SINE WAVE AMPLITUDE
ACCUMULATION 50000
CHANNELS 128

Figure 5.2-6: Information analysis of a DNA histogram. %CV = 5, G0-G1 = 30%, S = 40%, G2-M = 30%, and rate of DNA synthesis is constant.

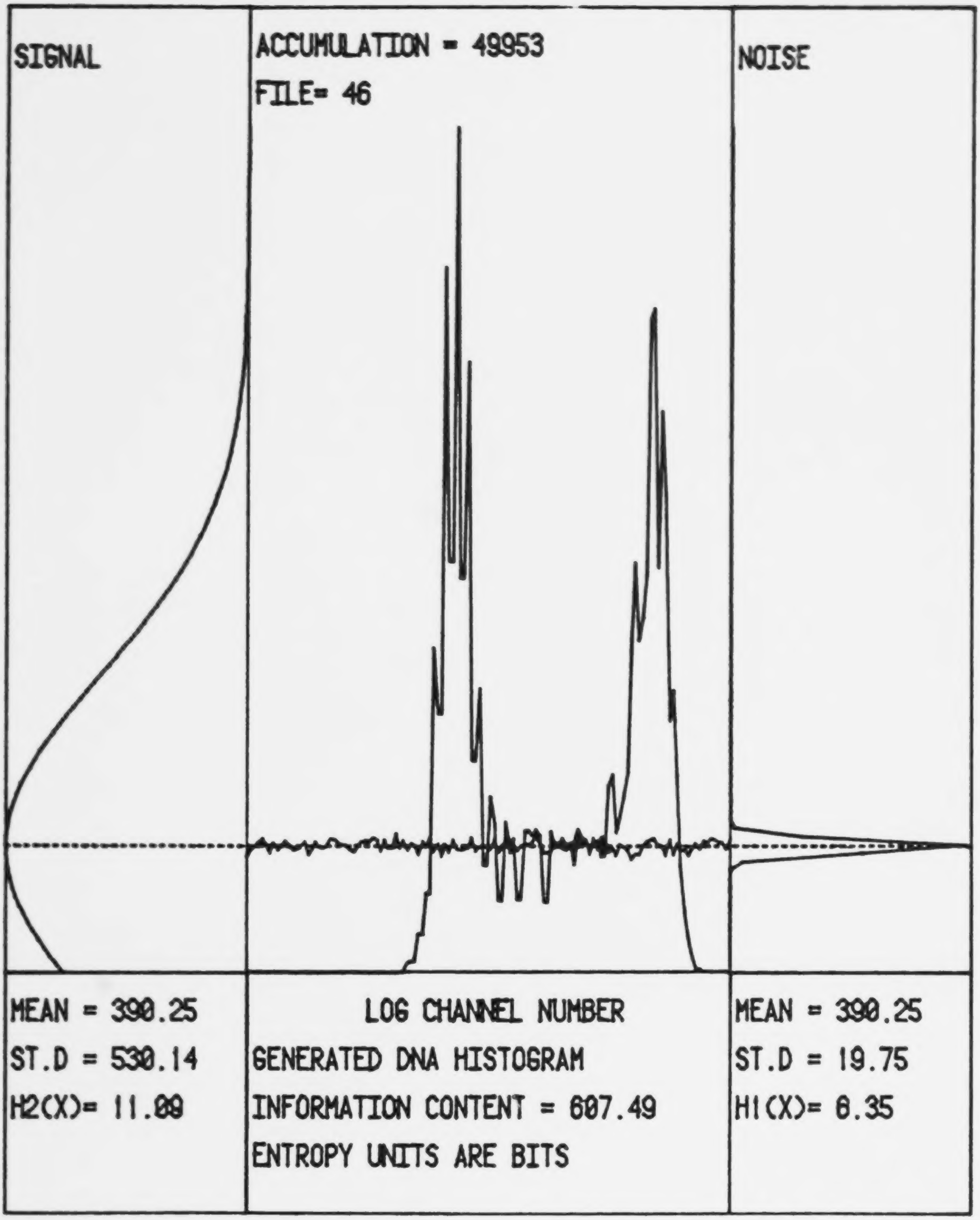
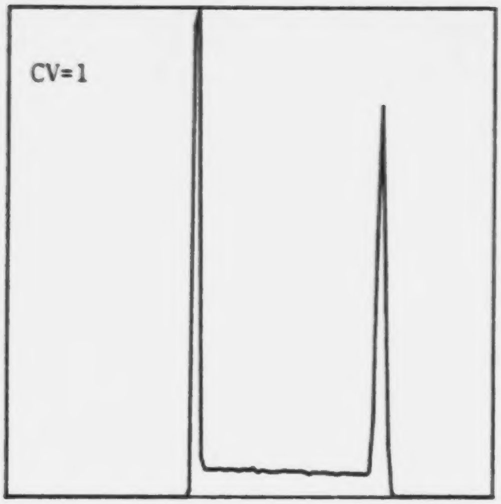
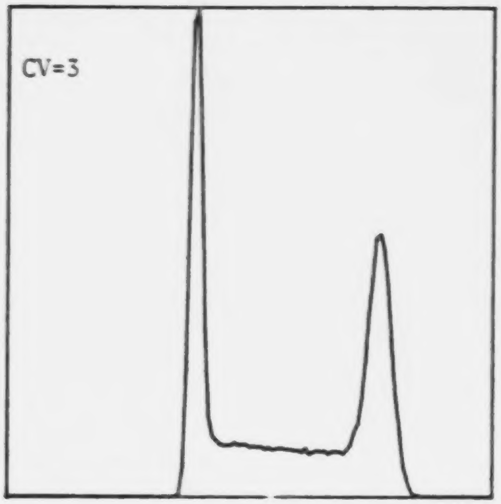

SIGNAL
ACCUMULATION = 49953
FILE= 46
NOISE
MEAN = 390.25
ST.D = 530.14
H2(X)= 11.09
LOG CHANNEL NUMBER
GENERATED DNA HISTOGRAM
INFORMATION CONTENT = 607.49
ENTROPY UNITS ARE BITS
MEAN = 390.25
ST.D = 19.75
H1(X)= 6.35

Figure 5.2-7: Four DNA histograms with %CV's of 1, 3, 5, and 7. G0-G1 = 30%, S = 40%, G2-M = 30%, and rate of DNA synthesis is constant.





NUMBER OF CELLS = 50,000

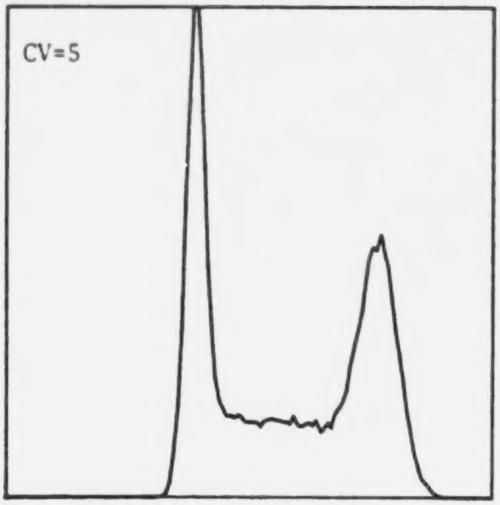
PERCENT G0-G1 = 30%

PERCENT S = 40%

PERCENT G2-M = 30%

G0-G1 MEAN = 50

G2-M MEAN = 100



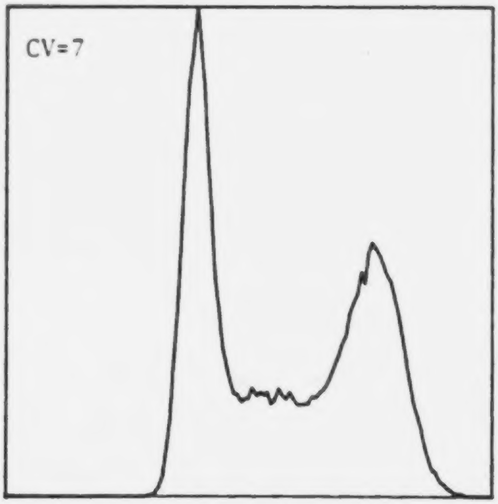
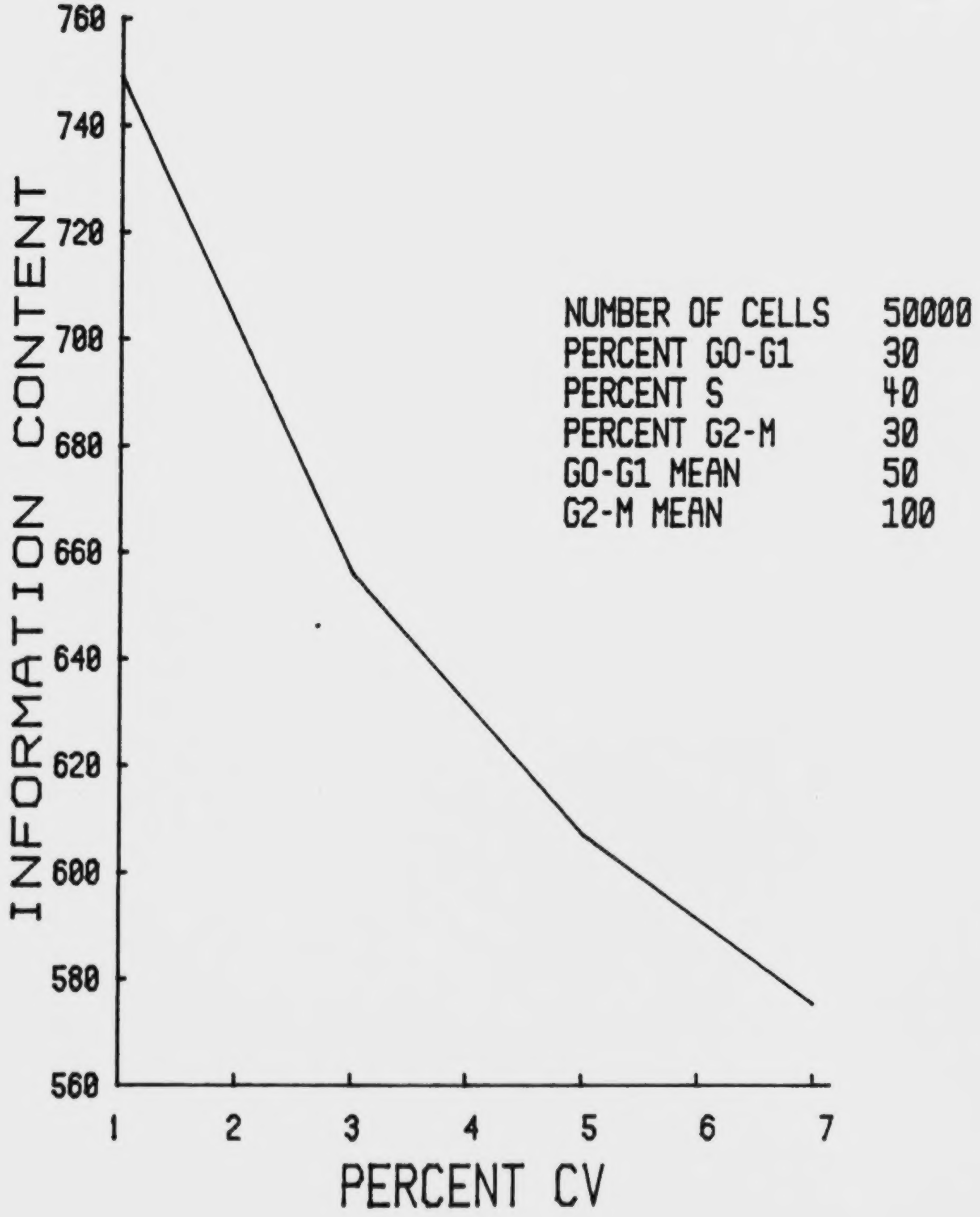


Figure 5.2-8: Information content vs %CV.


INFORMATION CONTENT
760
740
720
700
680
660
640
620
600
580
560
1
2
3
4
5
6
7
PERCENT CV
NUMBER OF CELLS 50000
PERCENT G0-G1 30
PERCENT S 40
PERCENT G2-M 30
G0-G1 MEAN 50
G2-M MEAN 100

5.3 Information density.

5.31 Line spread.

Suppose we have a single initial message that is transmitted across noisy space. The message is randomly perturbed during its travel and when it arrives at the receiver, it no longer is a discrete message but is now a probability distribution. A simple model of this process is "Galtons Quicunx". If marbles are started rolling at the top of a slanted pegged board, they will be randomly perturbed by the pegs and form a distribution at the bottom of the board (see Figure 5.31-1). The observed distribution can best be described by the binomial distribution. The binomial distribution requires lengthy calculations of large factorials and therefore is usually approximated by the Gaussian distribution. Figure 5.31-2 shows the observed frequency distribution, the binomial distribution, and the Gaussian distribution. The term "line spread" refers to this spreading effect of noisy space. If one has a constant line spread, then the standard deviation of the final accumulation will be the same regardless of where the message originates. For most flow cytometers, the line spread increases linearly with higher channel numbers. The ideal relationship between the standard deviation and channel number is exemplified in Figure 4.5-1. Figure 5.31-3 demonstrates the effect of a linear line spread on multiple messages.

Figure 5.31-1: Fundamental model of line spread of a single message.

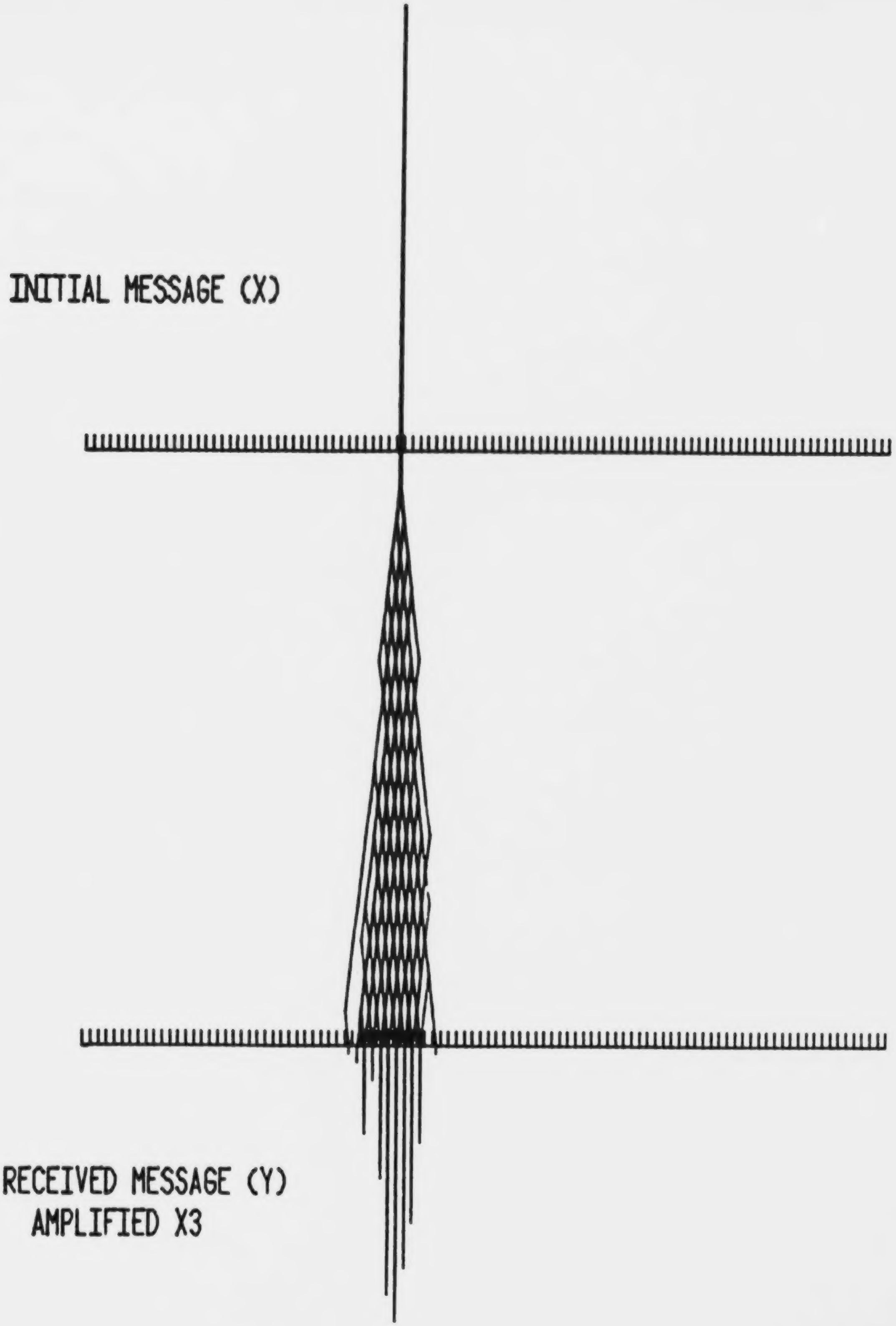

INITIAL MESSAGE (X)
RECEIVED MESSAGE (Y)
AMPLIFIED X3

Figure 5.31-2: Comparison of observed frequency histogram, binomial distribution, and Gaussian distribution.

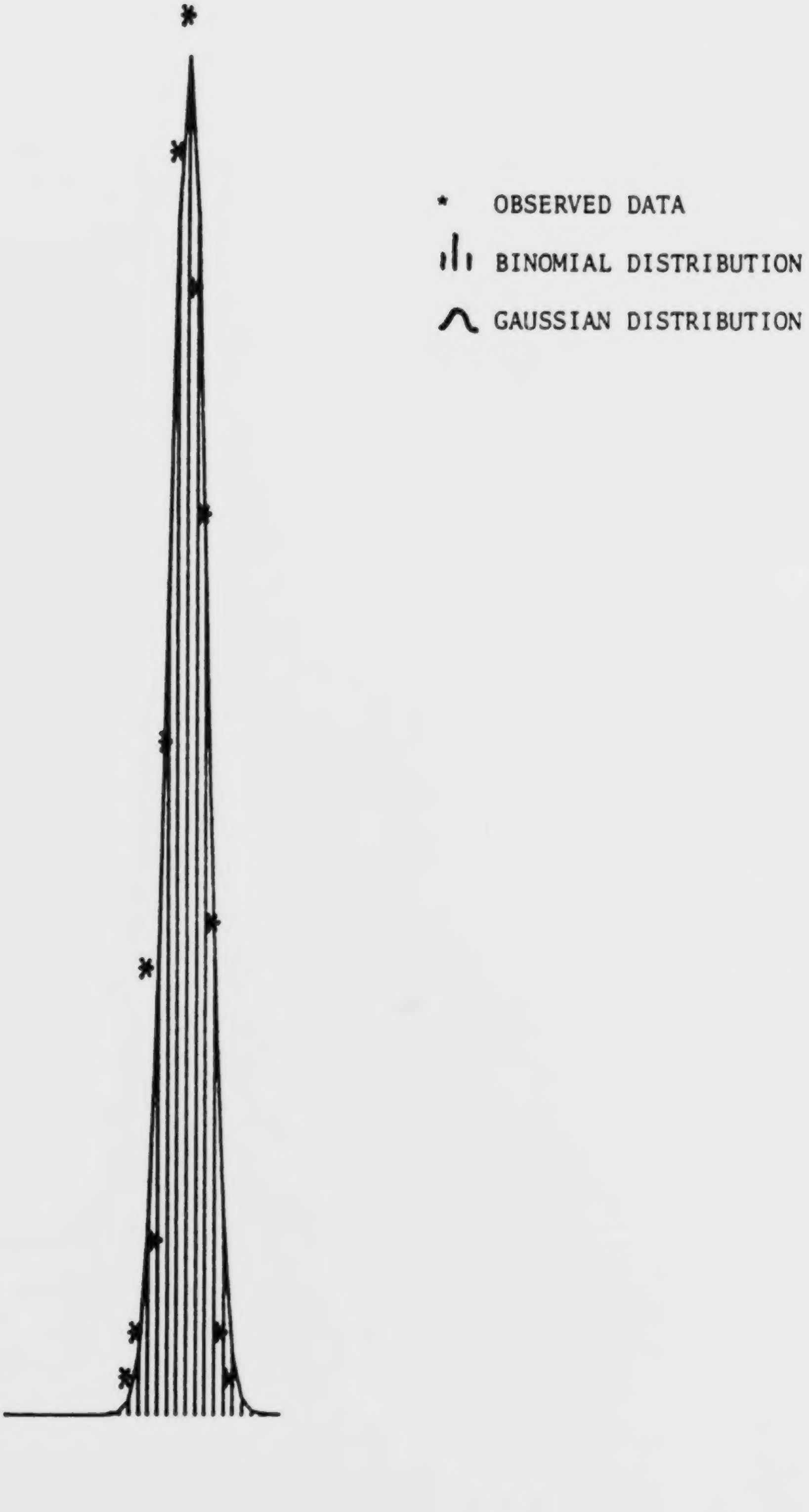
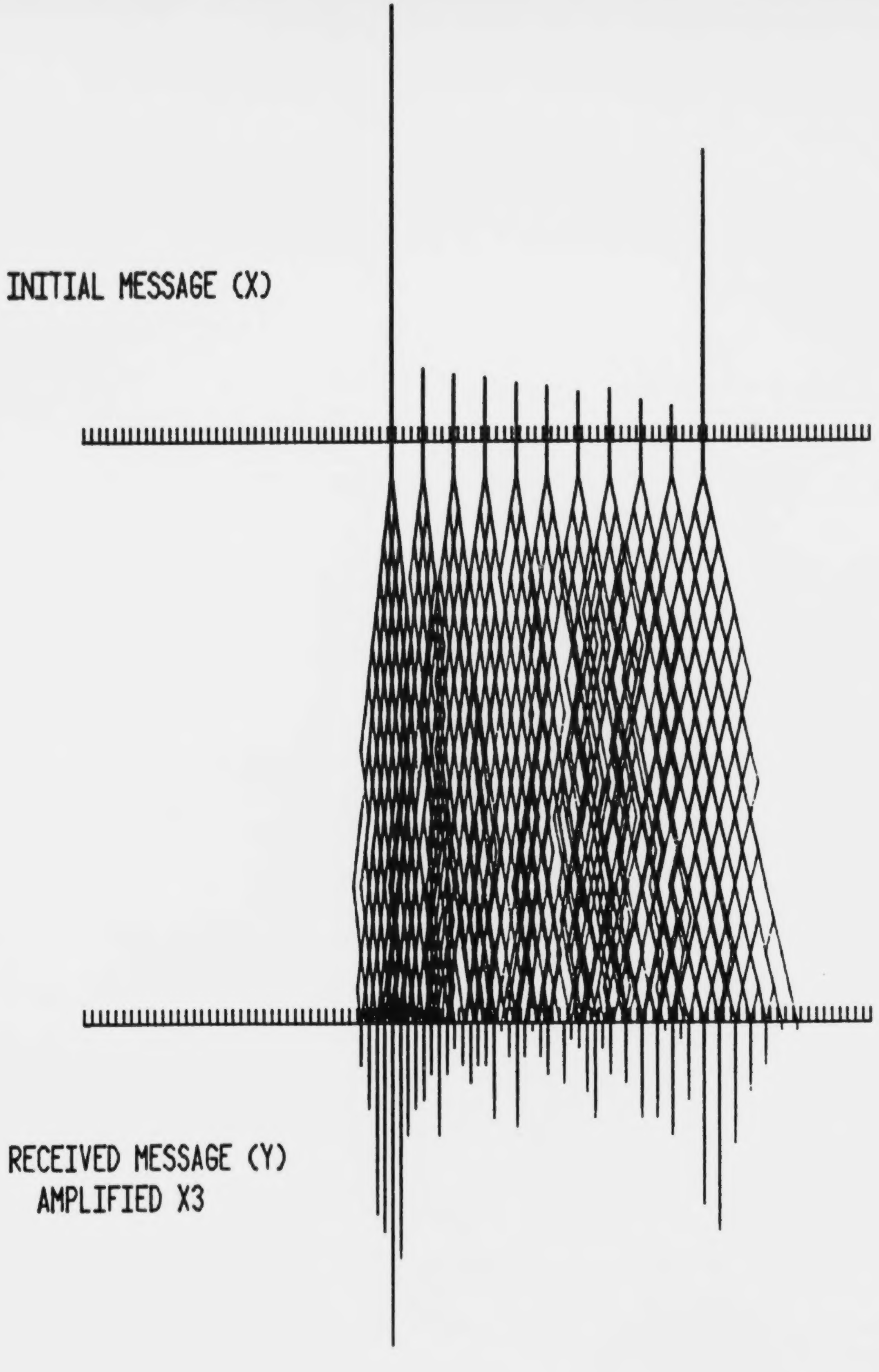

* OBSERVED DATA
ılı BINOMIAL DISTRIBUTION
⌒ GAUSSIAN DISTRIBUTION

Figure 5.31-3: Linear line spread with multiple messages.


INITIAL MESSAGE (X)
RECEIVED MESSAGE (Y)
AMPLIFIED X3

5.32 Information density of a single message.

Let us say we observe a Gaussian distribution on our flow cytometer. The number of counts is given by N and its standard deviation is found to be SD, and the mean is computed to be M. We know that M approximates the channel number where the initial message came from, but how good is this approximation? It can be shown (17) that the standard error of the mean (SEM) is equal to

$$SEM = \frac{SD}{SQR(N)} \qquad (5.32\text{-}1).$$

This formula states that our uncertainty of the actual mean is inversely proportional to the square root of the number of counts. Let us assume that we have a flow cytometer with constant line spread. Our original uncertainty of what the message was is simply

$$H(Xc) = LOG_2(C) \qquad (5.32\text{-}2)$$

where C equals the number of channels. Our uncertainty after having observed the distribution (Y) is given by equation 5.2-3, that is

$$H(Xc/Y) = .5 * LOG \left(\frac{2 * PI * e * SD^2}{N} \right) \qquad (5.32\text{-}3)$$

The amount of information we gain is our reduction in uncertainty after the observation Y or

$$Ic = H(Xc) - H(Xc/Y) = .5 * \mathrm{LOG}_2 \left(\frac{C^2 * N}{2*PI*e*SD^2} \right) \quad (5.32\text{-}4)$$

where Ic is the information gain for a constant line spread. Now, suppose we have a flow cytometer with a linear line spread (constant CV). As mentioned in the previous chapter, the probability for a message being found in any channel is not constant for this configuration. The probability of finding an accumulation in its correct channel is given by equation 5.2-5. If one normalizes the integration of this probability to unity over all channels, the result is

$$Pli = \frac{1}{LN(C)*i}$$

where C is the number of channels and i is the channel number. We can now determine the entropy in NITS of such a probability distribution by integrating the entropy over all channels.

$$H(X1) = \sum_{i=1}^{C} P1i * LN\left(\frac{1}{Pi} \right) \qquad (5.32\text{-}5)$$

$$= \frac{1}{2*LN(C)} * LN\left(LN(C)*i \right)^{2} \Big|_{1}^{C}$$

$$= \frac{1}{2*LN(C)} * \left[(LN(C*LN(C)))^{2} - (LN(LN(C)))^{2} \right]$$

We can simplify this equation by assuming that

$LN(C) \gg LN(LN(C))$

$$H(X1) = .5 * LN(C) \text{ NITS or}$$

$$= .5 * LOG_{2}(C) \text{ BITS} \qquad (5.32\text{-}6)$$

This equation indicates that the maximum information gain for the constant line spread configuration is twice or 200% the normal linear line spread configuration. The information gain for a single message observation is

$$Ili = H(X1) - H(X1/Y) = .5 * LOG_{2}\left(\frac{C * N}{2*PI*e*SDi^{2}} \right) \qquad (5.32\text{-}7)$$

The only difference between this equation and 5.32-4 is the squared term for C and the position dependent SDi.

5.33 Information density of multiple messages.

Now, suppose we have multiple messages like those shown in Figure 5.31-3 with linear line spread. The only difficulty in using equation 5.32-7 to estimate the information gain at a particular channel is that the number of accumulations associated with message i (Ni) is not exactly known. However, Ni can be estimated. If we knew the true peak height, PHi, of the underlying Gaussian distribution associated with the message that originated from channel i, we could calculate Ni by the formula

$$Ni = SQR(2*PI) * SDi * PHi \qquad (5.33\text{-}1)$$

which is simply the area of a Gaussian with a peak height of PHi and a standard deviation of SDi. Let us assume that the observed accumulation at channel i, Yi, is the summation of the underlying PHi and the contribution of channel i's nearest neighbor distributions (i-1 and i+1).

$$Yi = PHi+PH(i-1)*e^{-.5(1/SD(i-1))^2} + PH(i+1)*e^{-.5(1/SD(i+1))^2} \qquad (5.33\text{-}2)$$

If we make the assumption that PHi ~ PH(i+1) ~ PH(i-1) and that SDi ~ SD(i-1) ~ SD(i+1), then equation 5.33-2 simplifies to

$$Yi \sim PHi * (1 + 2*e^{-.5*(1/SDi)^2}) \qquad (5.33\text{-}3)$$

If we let Ki be defined as

$$Ki = \frac{1}{1 + 2*e^{-.5*(1/SDi)^{2}}} \qquad (5.33\text{-}4)$$

We can estimate Ni by

$$Ni \sim SQR(2*PI) * Ki * Yi \qquad (5.33\text{-}5).$$

We can plug equation 5.33-5 into equation 5.32-7 to determine the final form of the information density at channel i of a histogram with linear line spread

$$Ili = .5 * LOG_{2} \left(\frac{C * Ki * Yi}{SQR(2*PI)*e*SDi} \right) \qquad (5.33\text{-}6)$$

The information density for a constant line spread is similarly given by

$$Ici = .5 * LOG_{2} \left(\frac{C^{2} * K * Yi}{SQR(2*PI)*e*SD} \right) \qquad (5.33\text{-}7)$$

We can sum these information density equations over all the C channels to obtain the total information gain.

$$Ic = \sum_{i=1}^{C} Ici \qquad (5.33\text{-}8)$$

$$Il = \sum_{i=1}^{C} Ili \qquad (5.33\text{-}9)$$

In chapter 5.2 we found the expression for structural information. These equations were derived such that if the signal represented random noise, the calculated structural information would be 0. The equatins in this chapter; however, have an entirely different interpretation and application. They calculate the number of BITS of information gained in the process of observation. Another way of saying this is that they represent the number of BITS required to find any set of messages. A random set of messages will yield a positive information gain.

In chapter 5.2 we found that the structural information to be rather weakly influenced by CV. In fact, we showed an example where a reduction of CV by a factor of 5 (5 to 1) only increased the structural information by 21%. We can perform a similar computation with formula 5.33-9. Suppose we have 128 channels with 300 counts in each channel. We ask what is the information gain for a CV of 1 vs a CV of 5. The answer is 571 and 436 BITS respectively. The percent increase is only 26.8%. We can obtain an appreciation of how the total information gain is a function of CV and counts per channel (Yi) by inspection of the following table.

Table 5.32-1: Information gain vs %CV and counts per channel

		Counts per channel Yi			
		100	300	500	700
%CV	1	501	571	604	626
	3	405	475	508	529
	5	366	436	469	490
	7	341	411	444	466

I used the principles developed in this chapter to develop an optimal DNA histogram smoothing algorithm (Chapter 6.2) and a method for determining the optimal number and size of S phase compartments (Chapters 7.621 and 7.622). In the next chapters I will discuss the techniques of extracting this information from DNA histograms. The methods of extracting this information are basically divided into two catagories: parametric analysis and nonparametric analysis. In parametric analysis, we are attempting to find the most likely distribution of cell cycle phases that describe the data. This is done in three distinct levels of accuracy: crude stripping, pattern recognition adjustment, and multirectangle nonlinear least squares analysis (parametric analysis II). Nonparametric analysis is simply concerned with the localization of statistically significant areas within DNA histograms from two groups. This analysis will be applied later to the clinical problem of parathyroid disease (Chapter 9).

6. Parametric analysis I.

6.1 Introduction.

Parametric Analysis I is a system of subroutines that analyze DNA histograms. The analysis is parametric in the sense that the 128 numbers that constitute a DNA histogram are reduced to four biologically meaningful parameters (eg. %G0-G1, %S, %G2-M, and CV). All the routines can be linked together with the automatic analysis routine. A normal analysis processes the data as follows:

1. Smoothing
2. Capping
3. Stripping
4. Pattern Recognition Adjustment
5. Translocation
6. Doublet Correction

The final output is the mean position, standard deviation, and peak height of all the peaks in a DNA histogram. From these parameters the program automatically calculates the %G0-G1, %S, %G2-M, and CV.

6.2 Smoothing Routine.

6.21 Introduction.

Occasionally it is desirable to smooth the DNA histogram data. There are only two real advantages in smoothing. By dampening the statistical high frequency noise in a histogram, the peak searching and capping routines (next chapter) are much more reliable and accurate. The second advantage is that the statistical comparison of previously smoothed data (nonparametric analysis) is more sensitive. There have been a number of smoothing routines reported (18), but they all cause disturbingly high distortions in the structural information of the histogram. This distortion usually can be traced to the assumption that structural information is of low frequency. This is not the case with DNA histograms. The Fourier transform of an "ideal" DNA histogram contains very high frequency components to fit the sharp interphase boundaries between G0-G1 or G2-M and S. Thus, low frequency pass smoothing routines cannot reliably be used for DNA histograms.

The smoothing routine that is developed here has a variable window. The larger the window, the greater the smoothing. The size of the window is a function of the information density of the histogram. Thus, in areas of high information density (eg. G0-G1), this routine will have a small window. The window may be so small that there is actually no smoothing, but just an interpolation between observed points.

The heart of this smoothing routine is based on the simple observation that when Gaussians are plotted in the log domain (Y axis), they are parabolic. Therefore, a parabolic smoothing function was chosen because it matched the general contour of the DNA histogram in the log domain. The general operation of the smoothing routine can best be explained if the reader would imagine that he was a millimeter tall and walked out onto a sheet of paper containing a series of dots that constitutes a DNA histogram. Let us start at the first dot, that is, the accumulation in channel one. The reader is asked to draw a smooth line through all the dots. The reader has only one tool, a parabolic French rule. Let us say he looks ahead three points, slaps down his French rule, and draws the best parabolic arc through those points. He walks over to the end of his line, and repeats the process, but starting this time at the end of the drawn line. You can see that by looking ahead more points (increasing the window size) you are smoothing the data more. But when you look back, you see that at the intersections of the arcs, there are peaks and valleys. The problem is that the slope of the end of the preceding parabolic arc was not the same as the beginning slope of the next arc. Let us start over, but this time we will force the slopes at the intersection of two parabolic arcs to be the same. In fact, we can look ahead, let us say, five points, draw a small portion of the total arc to the next point, then repeat this process making sure that the slopes at the intersections are forced to be the same. If there were a voice in the background that would tell us how many points to look ahead each time, we would have a pretty good smoothing routine that would give reasonable results. This is exactly how the smoothing routine works. The voice in the background is the estimation of the information density. The mathematics for the most part are easy. The only complicated aspect was finding the optimal coefficients that related information density to window size.

6.22 Mathematics.

Let the smoothing function be described by the parabola

$$Y = C1*X^2 + C2*X + C3 \quad (6.2\text{-}1).$$

If we force the function to pass through Xo,Yo, then

$$Yo = C1*Xo^2 + C2*Xo + C3 \quad (6.2\text{-}2).$$

By subtracting eq 6.2-2 from eq 6.2-1, we obtain

$$Y-Yo = C1*(X^2 - Xo^2) + C2*(X - Xo) \quad (6.2\text{-}3).$$

Let us further restrict this function such that it must begin with slope Bo. Bo is equivalent to the first derivative of eq 6.2-2

$$Bo = 2*C1*Xo + C2 \quad (6.2\text{-}4).$$

If we solve for C2 in equation 6.2-4 and substitute it into equation 6.2-3, we obtain

$$Y-Yo = C1*(X-Xo)^2 + Bo*(X-Xo) \quad (6.2\text{-}5).$$

If we let N equal the number of points the smoothing routine looks ahead and Yi is the log of the observed accumulation in channel i, then the least squares expression would be

$$SSQ = \sum_{i=1}^{N} \left[Yi - Yo - C1*(Xi-Xo)^2 - Bo*(Xi-Xo) \right]^2 \quad (6.2\text{-}6)$$

By taking the derivative of SSQ with respect to C1 and setting it equal to 0, one obtains

$$(Yi-Yo)*(Xi-Xo)^2 = C1*(Xi-Xo)^4 + Bo*(Xi-Xo)^3 \quad (6.2\text{-}7)$$

Solving for C1, one obtains the formula

$$C1 = \frac{(Yi-Yo)*(Xi-Xo)^2 - Bo*(Xi-Xo)^3}{(Xi-Xo)^4} \quad (6.2\text{-}8)$$

where C2 is simply given by

$$C2 = Bo - 2*C1*Xo \quad (6.2\text{-}9).$$

C3 is calculated by

$$C3 = Yo - C1*Xo^2 - C2*Xo \quad (6.2\text{-}10).$$

The next slope in the smoothing routine would be

$$B1 = 2*C1*X1 + C2 \quad (6.2\text{-}11).$$

The next Y coordinate would be

$$Y1 = C1*X1^2 + C2*X1 + C3 \quad (6.2\text{-}12).$$

This process continues throughout the entire DNA histogram. For each point on the DNA histogram, we calculate a C1, C2, and C3 that best describes a parabola passing through the next N points. It should be noted that the Y variable in the above formulation is actually the log of the number of cells observed in each channel and N is a function of the information density at Xi.

6.23 Variable window size and optimization.

The information density at a given channel is going to be a function of the accumulation number, Yi, and the computed line spread standard deviation (SDi; see equation 5.33-6). We make the assumption that the optimal window width, N, is linearly related to the information density at channel i, that is

$$N = S1 - S2 * \ln\left(\frac{Yi}{SDi} \right) \quad (6.2\text{-}13)$$

By a trial and error type of process, a better function involved the averaging of Yi with the next observed value. Also, it was found by smoothing from left to right we avoid undesirable steep walls associated with low channel gating. Thus, the final function used was

$$N = S1 - S2 * \ln\left(\frac{Yi\text{-}1 + Yi}{SDi} \right) \quad (6.2\text{-}14).$$

The process of calculating S1 and S2 starts by establishing a test case. The test case finally decided on was 70 Gaussians of increasing standard deviation (0.5-5) with 7 levels of peak heights (10, 32, 100, 316, 1000, 3162, and 10000); see Figure 6.2-1). One hundred and thirty-three sample areas were selected (all peak and trough points) and stored in an array, [N-]. Figure 6.2-1 marks the location of these points by X's and Figure 6.2-2 shows the actual array contents. Statistical noise was added to the histograms such that the SD was equal to the square root of the noiseless value. The function (19) that was used to generate these noisy Gaussians is

Yr = SQR(-2*ln(RND))*COS(2*PI*RND)*SQR(Y) + Y (6.2-15).

where Yr is the random Y value and RND is a random number generator ($0 \le RND \le 1$). A noisy array, [N+], was obtained in the same manner as the [N-] array. The overall contribution of noise to this system can best be described by adding the square of the differences between arrays [N-] and [N+]. Thus, we have

$$[SC] = \left[[N-] - [N+] \right]^2 \qquad (6.2\text{-}16)$$

$$SSC = \sum_{ij} [SC]$$

Let us now make reasonable guesses as to what S1 and S2 are (eg. 8, 2). We smooth the noisy set of Gaussians with these given S1 and S2 values and arrive at a third characteristic array, [NS]. We can obtain an idea of how well we smoothed by looking at the difference array, [DIFF], defined as

$$[DIFF] = \left| [N-]-[N+] \right| - \left| [N-]-[NS] \right|^{2}$$

If any of the elements of this array are positive, then the smoothing routine resulted in a better estimation of the true noiseless value at this point. Negative elements are where the smoothing routine biased the data to the point where it was hurting rather than helping. Zero elements would, of course, be the break-even point. If all the elements were summed in [DIFF] the result would be an overall indication of how well we smoothed (SDIFF). The higher the number, the better the smoothing routine. The values of S1 and S2 were incremented in a grid search to find the maximum SDIFF value. In order to guard against optimizing the S1 and S2 values for this particular noisy test case, random noise was introduced with every iteration. At the end of this search routine, the optimized [DIFF] was obtained. There were definite regions where, even when optimized, the elements were negative (eg. low SD and high counts). A smoothing boundary was chosen on the basis of these results that would signify where the smoothing routine had the best chance of yielding results closer to the noiseless values. If the routine was in a negative region it would not smooth, but just interpolate between points. The optimization program was run again with this boundary constraint, and the final values of S1 and S2 were found to be 6.2 and 2.4 respectively (see Figure 6.2-3).

A typical log DNA histogram, raw and smoothed, is shown in Figure 6.2-4. The smoothing characteristics of this algorithm are quite mild and have not yielded unreasonable results in the five to ten thousand histograms that have been smoothed in this manner.

Figure 6.2-1: Noiseless seventy Gaussian test case for smoothing optimization.

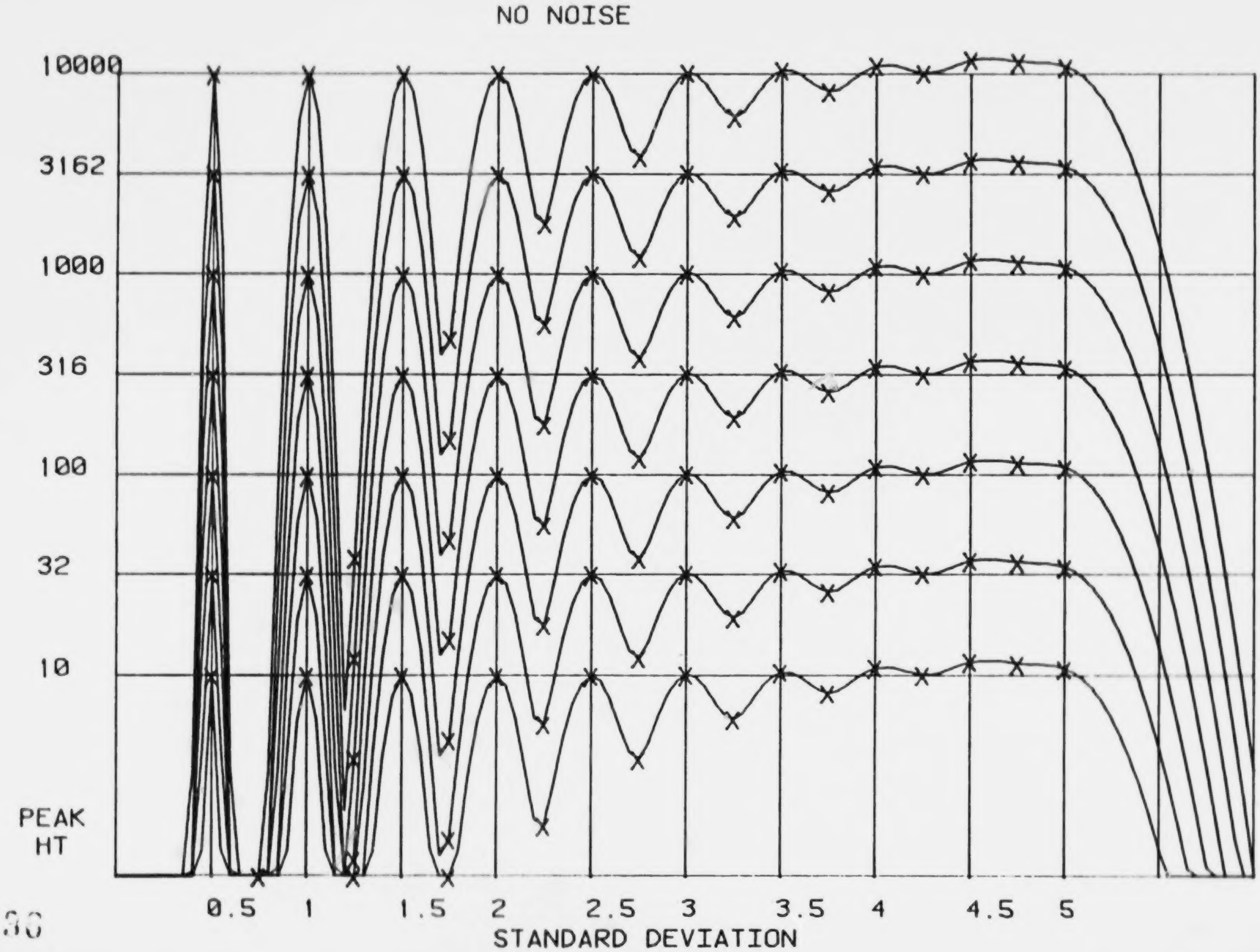

NO NOISE
10000
3162
1000
316
100
32
10
PEAK
HT
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
STANDARD DEVIATION

Figure 6.2-2: Noiseless test array [N-].

NO NOISE

PEAK HEIGHTS SUM= 214021.371083

ST. DEV.

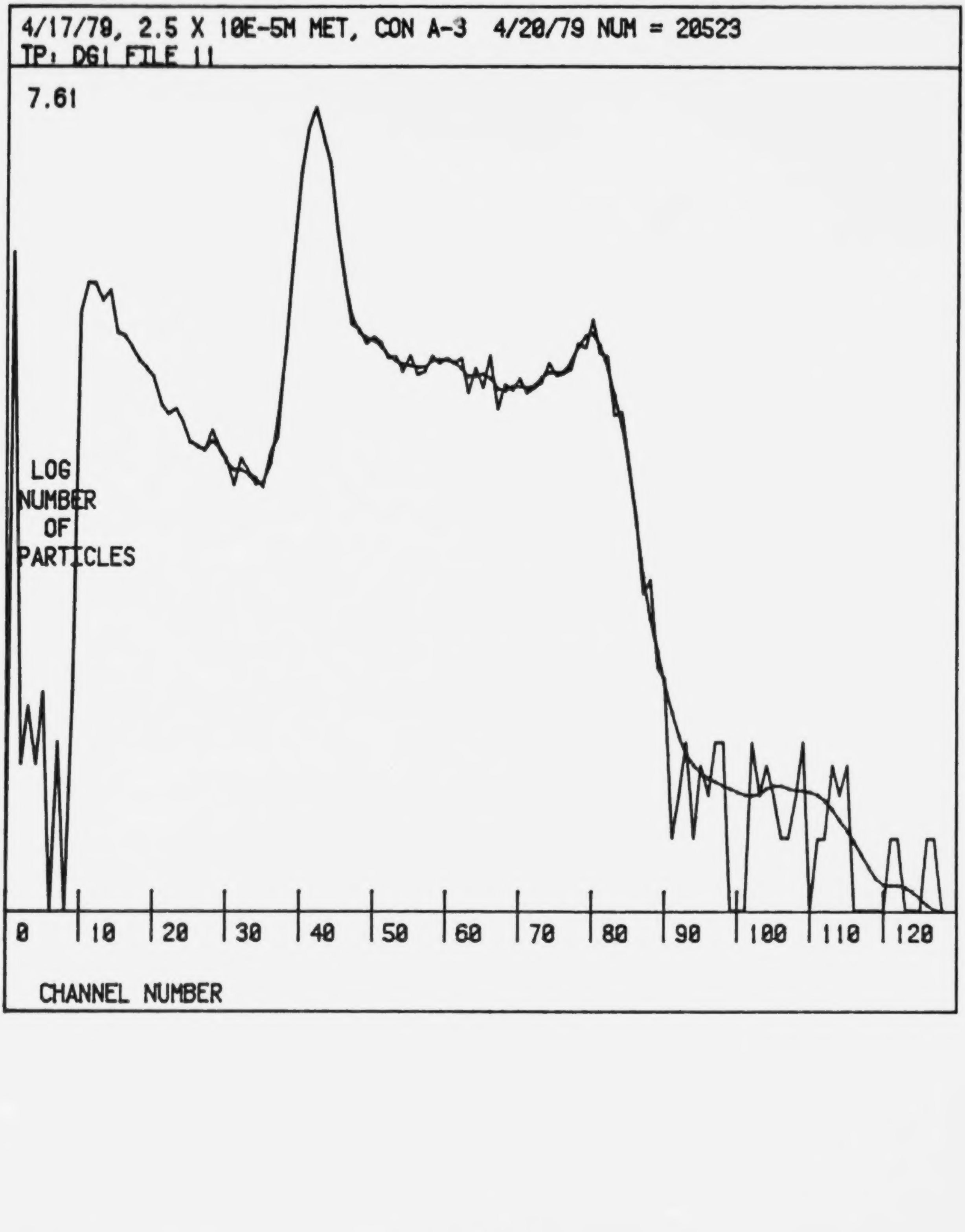
10	32	100	316	1000	3162	10000
0	0	0	0	0	0.01	0.03
10	32	100	316	1000	3162	10000
0.03	0.12	0.38	1.22	3.86	12.23	38.69
10	32	100	316	1000	3162.01	10000.03
0.47	1.52	4.78	15.1	47.8	151.15	478.02
10	32.01	100.03	316.1	1000.33	3163.06	10003.35
1.79	5.73	17.92	56.65	179.27	566.87	1792.75
10.03	32.12	100.38	317.22	1003.86	3174.23	10038.69
3.84	12.31	38.47	121.59	384.79	1216.7	3847.9
10.17	32.55	101.72	321.44	1017.21	3216.44	10172.19
6.1	19.54	61.06	192.97	610.68	1930.98	6106.83
10.47	33.53	104.78	331.12	1047.85	3313.31	10478.54
8.22	26.3	82.21	259.8	822.15	2599.65	8221.54
11.01	35.25	110.18	348.19	1101.87	3484.12	11018.73
10.08	32.27	100.84	318.67	1008.45	3188.72	10084.52
11.79	37.73	117.92	372.65	1179.27	3728.85	11792.72
11.46	36.69	114.68	362.39	1146.82	3626.25	11468.21
10.84	34.7	108.46	342.75	1084.66	3429.7	10846.61

Figure 6.2-3: Optimized [DIFF] array. Positive elements indicate where smoothing yielded better results than noisy values alone.

```
                     OPTIMIZED [DIFF] ARRAY

          <------------------- PEAK HEIGHTS ------------------->
--------------------------------------------------------------------
1 STD DEV 1 10       32      100     320     1000    3200   10000      1
--------------------------------------------------------------------1
1   .5    1  0        0       0       0       0       0       0        1
1         1                                                            1
1         1  0        0       0       0       0       0       0        1
1         1                                                            1
1  1.0    1  0        0       0       0       0       0       0        1
1         1                                                            1
1         1  0        0       0       0       0       0       0        1
1         1                                                            1
1  1.5    1  0        0       0       0       0       0       0        1
1         1-------------                                               1
1         1  1        0  1    0       0       0       0       0        1
1         1              1                                             1
1  2.0    1 -4       16  1    0       0       0       0       0        1
1         1              1                                             1
1         1  9        9  1    0       0       0       0       0        1
1         1              1                                             1
1  2.5    1  10      60  1    0       0       0       0       0        1
1         1              1                                             1
1         1  1        9  1    0       0       0       0       0        1
1         1              -------------------------1                    1
1  3.0    1  4        1       63      2      -52  1  0       0         1
1         1                                       1                    1
1         1  13      33       350     263     520 1  0       0         1
1         1                                       1                    1
1  3.5    1 -5        6       45      277     167 1  0       0         1
1         1                                       1                    1
1         1  10      100      37      365    1890 1  0       0         1
1         1                                       1                    1
1  4.0    1  5       138      176     946   -1041 1  0       0         1
1         1                                       1                    1
1         1  50      10       22      6       635 1  0       0         1
1         1                                       ---------------------1
1  4.5    1  10       2       54      624    -127   6919     0         1
1         1                                                            1
1         1  -3      25      -49      38     4341  -1226     0         1
1         1                                                            1
1  5.0    1  3      -53       157    -14      897   8375   -2021       1
1         1                                                            1
--------------------------------------------------------------------
```

Figure 6.2-4: Smoothed and raw log DNA histogram.


4/17/79, 2.5 X 10E-5M MET, CON A-3 4/20/79 NUM = 20523
TP: DG1 FILE 11
7.61
LOG NUMBER OF PARTICLES
0
10
20
30
40
50
60
70
80
90
100
110
120
CHANNEL NUMBER

6.3 Capping Routine.

6.31 Introduction.

After the DNA histogram has been smoothed, it is important to know exactly where the first and last peaks of the histogram reside. This information is necessary for later translocation techniques. The capping routine, like the smoothing routine, is done in log space. The reason for the log transformation is because the peak of the DNA histogram has a parabolic shape in log space. The capping routine finds the point that represents the peak height of the first peak and then looks ahead and back a user defined number of standard deviations to find the best parabola through the points. Once the coefficients of this capping parabola have been determined, the derivative is taken and set to zero. The X Y coordinates that solve this first derivative equation are the new peak height and peak position of the first peak. If there is a second peak in the G2-M region, then capping is employed to find its peak height and position.

The capping routine, therefore, allows the program to extrapolate between channels and thereby give a more precise estimate to the location of the first and last peaks in the DNA histogram. If the histogram is to be doublet corrected

(Chapter 6.7), then capping is also used on the triplet peak (6C).

6.32 Mathematics.

If we let Y' be the log of the number of cells in each channel and X1 be the channel squared and X2 is the channel number, then the parabola that will cap either the first, second, or third peak is given by

$$Y' = C1*X1 + C2*X2 + C3 \quad (6.3\text{-}1).$$

We can simplify equation 6.3-1 if we first state that the mean of Y' (Y'm) over some region is equal to C1 times the mean of the channel numbers squared (X1m) plus C2 times the mean of the channel numbers (X2m) plus C3, thus

$$Y'm = C1*X1m + C2*X2m + C3 \quad (6.3\text{-}2).$$

If we subtract equation 6.3-2 from 6.3-1, we can cancel out C3.

$$Y'-Y'm = C1*(X1-X1m) + C2*(X2-X2m) \quad (6.3\text{-}3).$$

We are now ready for a least squares analysis. Let n be the number of observed points (Yi).

$$SSQ = \sum_{i=1}^{n} \left| Yi - Ym - (C1*(X1i - X1m) + C2*(X2i - X2m)) \right|^{2} \quad (6.3\text{-}4)$$

In order to simplify the math, the following substitutions are made:

$$[Y1] = \sum_{i=1}^{n} (Y'i-Y'm) * (X1i-X1m)$$

$$[12] = \sum_{i=1}^{n} (X1i-X1m) * (X2i-X2m)$$

and [11], [22], [Y2] are similarly defined.

If the partial derivative of SSC is taken with respect to C1 and C2, the following two equations are obtained

$$[Y1] - C1*[11] - C2*[12] = 0 \qquad (6.3\text{-}5)$$

$$[Y2] - C1*[12] - C2*[22] = 0 \qquad (6.3\text{-}6).$$

If we rearrange equations 6.3-5 and 6.3-6, then

$$[Y1] = C1*[11] + C2*[12] \qquad (6.3\text{-}7$$

$$[Y2] = C1*[12] + C2*[22] \qquad (6.3\text{-}8).$$

If we solve for C1 by multiplying equation 6.3-7 by [22] and 6.3-8 by [12] and subtracting the two equations, we obtain

$$C1 = \frac{[Y1][22] - [Y2][12]}{[11][22] - [12][12]} \qquad (6.3\text{-}9).$$

If we solve for C2 in the same manner we get

$$C2 = \frac{[Y1][12] - [Y2][11]}{[12][12] - [22][11]} \qquad (6.3\text{-}10).$$

C3 is simply given by the formula

$$C3 = Y'm - C1*X1m - C2*X2m \qquad (6.3\text{-}11)$$

If we take the derivative of equation 6.3-1 and set it to 0, we obtain

$$0 = 2 * C1*X1 + C2 \qquad (6.3\text{-}12).$$

The peak position and peak height of this capping parabola are

$$Xp = \frac{-C2}{2*C1} \quad \text{and} \quad Yh = C1*(Xp)^{2} + C2*Xp + C3 \qquad (6.3\text{-}13).$$

These X Y coordinates should represent a much better approximation of the true peak height and peak position of the first (and last) peaks in the DNA histogram.

6.4 Stripping Routine.

6.41 Introduction.

In the stripping routine we obtain our first estimates of the parameters that describe the major Gaussians in the DNA histogram. Since a histogram can have at least one, two, or three distinct peaks, the program provides the user with four options: 1) Search for only the first peak, 2) Search for only the G0-G1 peak and assume the peak position and standard deviation of the G2-M peak is exactly twice the G0-G1 peak. 3) Search for both the G0-G1 peak and the G2-M peak. 4) Search for the G0-G1 peak, the G2-M peak, and a triplet peak. In order to do a doublet correction (Chapter 6.7), the computer must know how many triplets are present.

The search technique is straightforward. The program requests only the boundaries that confine the first peak. The program finds the integer location of the first peak and then caps that peak. The capping boundaries are internally selected to be above and below a user-defined number of standard deviations away from the integer peak location. The capping routine calculates the peak position and height of the curve. The standard deviation of the curve is found in a different manner. The computer marches in both directions from the peak and determines which side of the peak has the steepest slope. It does this by comparing the half max values of the peak. Once it determines the side of steepest incline, it performs a least squares analysis between the peak and the half max position. The peak position and height are held constant and the only free parameter is the standard deviation.

After the standard deviation is calculated, the Gaussian is subtracted from the parent curve and the results are placed in a temporary register. If the user-selected option 2, the G2-M parameters are calculated directly from the G0-G1 parameters. If the user selected option 3 or 4, the G2-M parameters are calculated in the same way as the G0-G1 parameters were calculated. With option 4, the computer looks for a third peak representing triplets.

Once all the Gaussians have been "stripped" from the curve, S is calculated by adding up the residue from the first peak position to the second. The percentages of each compartment are determined and are outputed under the heading of "Crude Stripping." The accuracy of this algorithm will be discussed in chapter 6.5.

6.42 Mathematics.

Most of the mathematics associated with this routine are trivial. The only section of this routine that will be discussed is the mathematics associated with the least squares analysis of a Gaussian with a restricted peak height and position.

Let the peak position (Xp) and the peak height (Yh) of a Gaussian be given. The general equation of the Gaussian with the single free parameter SD is

$$Y = Yh * e^{-.5((X-Xp)/SD)^2} \quad (6.4\text{-}1)$$

Equation 6.4-1 can be converted to a parabola by taking the natural log of both sides

$$\ln Y = -\frac{(X-Xp)^2}{2*SD^2} + \ln Yh \quad (6.4\text{-}2).$$

If we let

$Y' = \ln y$

$X1 = (X-Xp)^2$

$C1 = \dfrac{-1}{2*SD^2}$

$Yh' = \ln Yh$ then,

$Y' = C1*X1 + Yh'$ (6.4-3).

Let SSQ be the summation of the squares of deviations of the observed Y values (Yi) from the calculated Y values.

$$SSQ = \sum_{i=1}^{n} \left| Yi - C1*X1i - Yh' \right|^2 \quad (6.4\text{-}4)$$

By taking the first derivative of SSQ and setting it equal to 0 we obtain

$$\sum_{i=1}^{n} Yi * X1i = C1 * \sum_{i=1}^{n} X1i * X1i + \sum_{i=1}^{n} Yh' * X1i \qquad (6.4\text{-}5)$$

Solving for C1,

$$C1 = \frac{\displaystyle\sum_{i=1}^{n} Yi * X1i - Yh' * \sum_{i=1}^{n} X1i}{\displaystyle\sum_{i=1}^{n} X1i^{2}} \qquad (6.4\text{-}6)$$

Therefore,

$$SD = SQR\left(\frac{-1}{2 * C1} \right) \qquad (6.4\text{-}7)$$

for $C1 < 0$ and if $C1 \geq 0$ then SD is set to 0.

6.5 Pattern Recognition Adjustment.

6.51 Introduction.

The main problem associated with DNA histogram analysis is the determination of the parameters that describe the G0-G1 and the G2-M Gaussian distributions. The reason why the solution to this problem is not trivial is that the S phase considerably overlaps into the G0-G1 and G2-M range when the CV is 4% or greater. There are a number of techniques available today that can accurately dissect the various cell cycle phases from each other. However, all these techniques involve initial guesses of the parameters and the final solution often depends quite heavily on the accuracy of the guesses.

The stripping routine tries to minimize the effects of S phase contamination by looking only at the side of the curve that is furthest from S. However, as the percentage of cells in S increase, the amount of error associated with the stripping routine also increases. The direction of error is always the same. The real position of G0-G1 is always to the left of the calculated position. As the percentage of cells in S increase, the distance between real and calculated position also increases. The peak height and standard deviation also appear to be greater than they actually are when there is significant S phase contamination. Figure 6.5-1 shows a reconstruction of a DNA histogram from its component parts and the errors associated with the approximation.

The degree of error made in estimating either the G0-G1 or G2-M parameters is also a function of CV. Figure 6.5-2 shows a histogram with the identical cell cycle parameters except the CV is 3% instead of 6%. Notice the reduction in the errors. It is for this reason that there is a great deal of effort being exerted in finding intercalating dyes or combination of dyes that yield low CV's.

The errors in estimation are mainly a function of the amount of S phase present and the CV of the distribution. It should also be mentioned that the shape of S phase, near G0-G1 is also important in contributing to the error of estimation. Now suppose we had a large library of generated histograms representing all the possible combinations of %G0-G1, %S and %G2-M and we made these libraries for a CV of 1% through 12% at 1% intervals. Let us say we catalogued the errors made with a particular CV and amount of S in some retrievable format. If we then were confronted with an unknown histogram we could determine its approximate CV and quantity of S and find the magnitude and direction of all the errors associated with the first estimate. We could then adjust the initial estimates appropriately and obtain a more accurate The accuracy of the results, however, will be influenced by deviations from the assumed exponential shape of S.

This process is called pattern recognition adjustment. The CV and quantity of S are called features of the pattern space. Actually, we use the relative peak height of S at some arbitrary point near G0-G1 as our feature for the quantity of S. Figure 6.5-3 shows a family of curves with different CV's relating the peak height estimation error with the relative peak height of S.

This process works relatively well with the estimation of G0-G1; however, it did not work for G2-M. The obvious trend that is shown in Figure 6.5-3 was absent. The degree of overlap between G2-M and S is so great that it may be impossible to employ this type of refinement to the G2-M parameters.

6.52 Mathematics.

The computer determines the correct error factor, F, by performing a number of weighted averages. In this example, we will be only concerned about the peak height error factor. Let us say that we determine the relative peak height of S (peak height of S divided by the 1st modal height X 100) in some DNA histogram to be 32% and the CV of the G0-G1 peak to be 6.5%. The computer calculates the error factor in the following manner.

1. First it determines the error factors recorded on tape on either side of 32 for a CV of 6%. Let us say they were .85 for a relative peak height of 28.5% and .81 for a relative peak height of 33.5%.

2. The program then does a weighted average to determine the best error factor for a relative peak height of 32%.

$$F(32\%) = .85 - \left\{(.85-.80)(33.5-32)/(33.5-28.5)\right\} = .835$$

3. The program then calculates the best error factor for a relative peak height of 32% with the CV=7% library. Let us say that it turned out to be .862.

4. The program finds the weighted average for the two error factors at a CV of 6.5%.

$$F(32\%,6.5\%) = .862 - \{(.862-.835)(7-6.5)\} = .845$$

5. Finally, the program multiplies the true peak height by the error factor to determine the adjusted peak height.

This process is repeated for the G0-G1 standard deviation and peak position. Each of these parameters has its own set of error factors associated with the relative peak height of S. Figure 6.5-3 summarizes this process for G0-G1 peak height adjustment. Figures 6.5-4A and B compare the actual percent S with the calculated percent S with and without pattern recognition. Most of the error is attributed to the inability to perform pattern recognition on G2-M.

Figure 6.5-1: DNA histogram and its components with a CV of 6%, G0-G1=30%, S=50%, G2-M=20%. Errors associated with assuming 1) G0-G1 peak height equals the 1st modal peak height. 2) G0-G1 peak position equals the 1st modal peak position. 3) G0-G1 standard deviation equals the 1st peak's standard deviation. 4) G2-M peak height equals the 2nd modal peak height. 5) G2-M peak position equals the 2nd modal peak position. 6) G2-M standard deviation equals the 2nd peak's standard deviation.

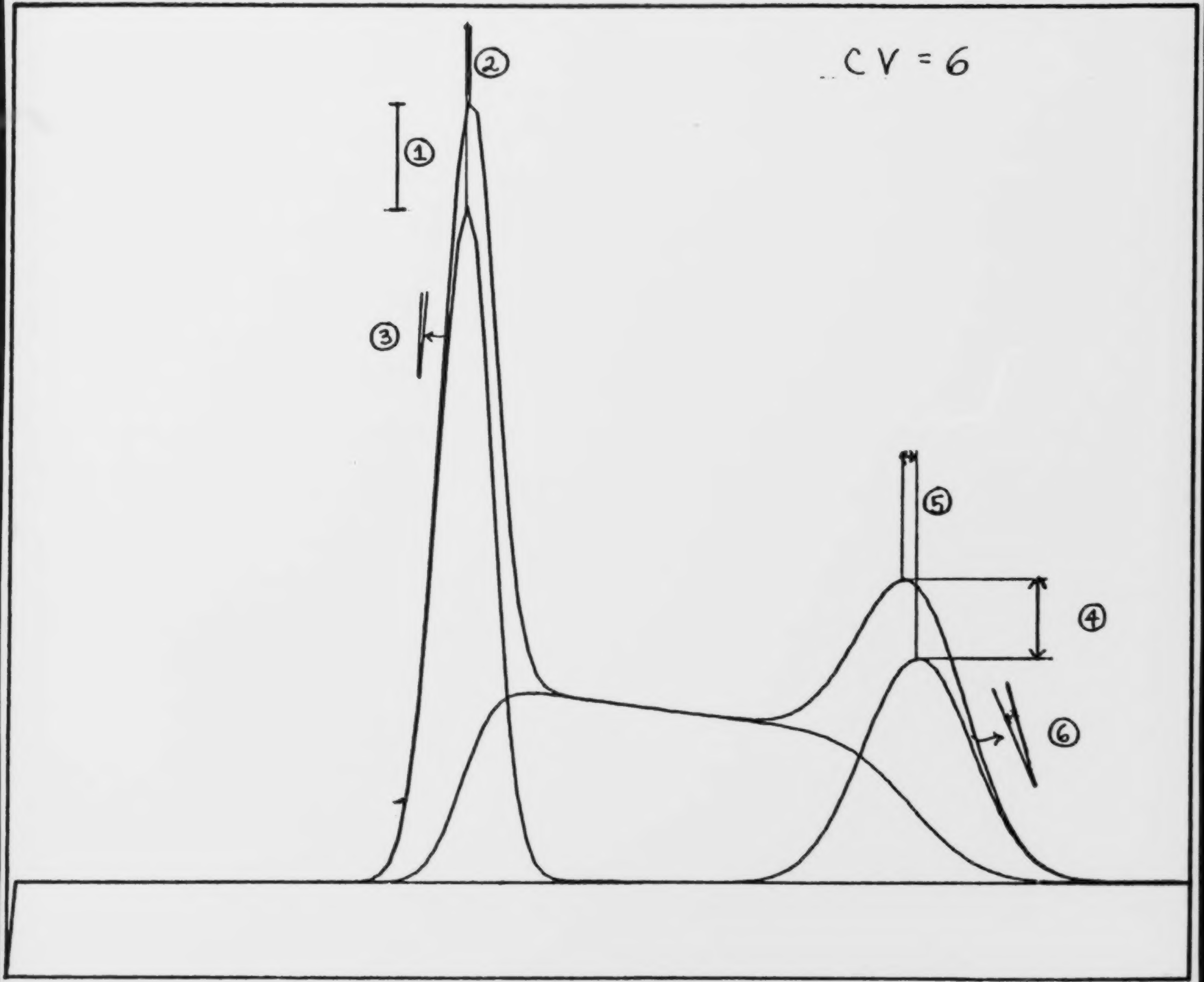

CV = 6
②
①
③
⑤
④
⑥

Figure 6.5-2: DNA histogram and its components with a CV of 3%. This figure has the same compartments and errors as Figure 6.5-1.

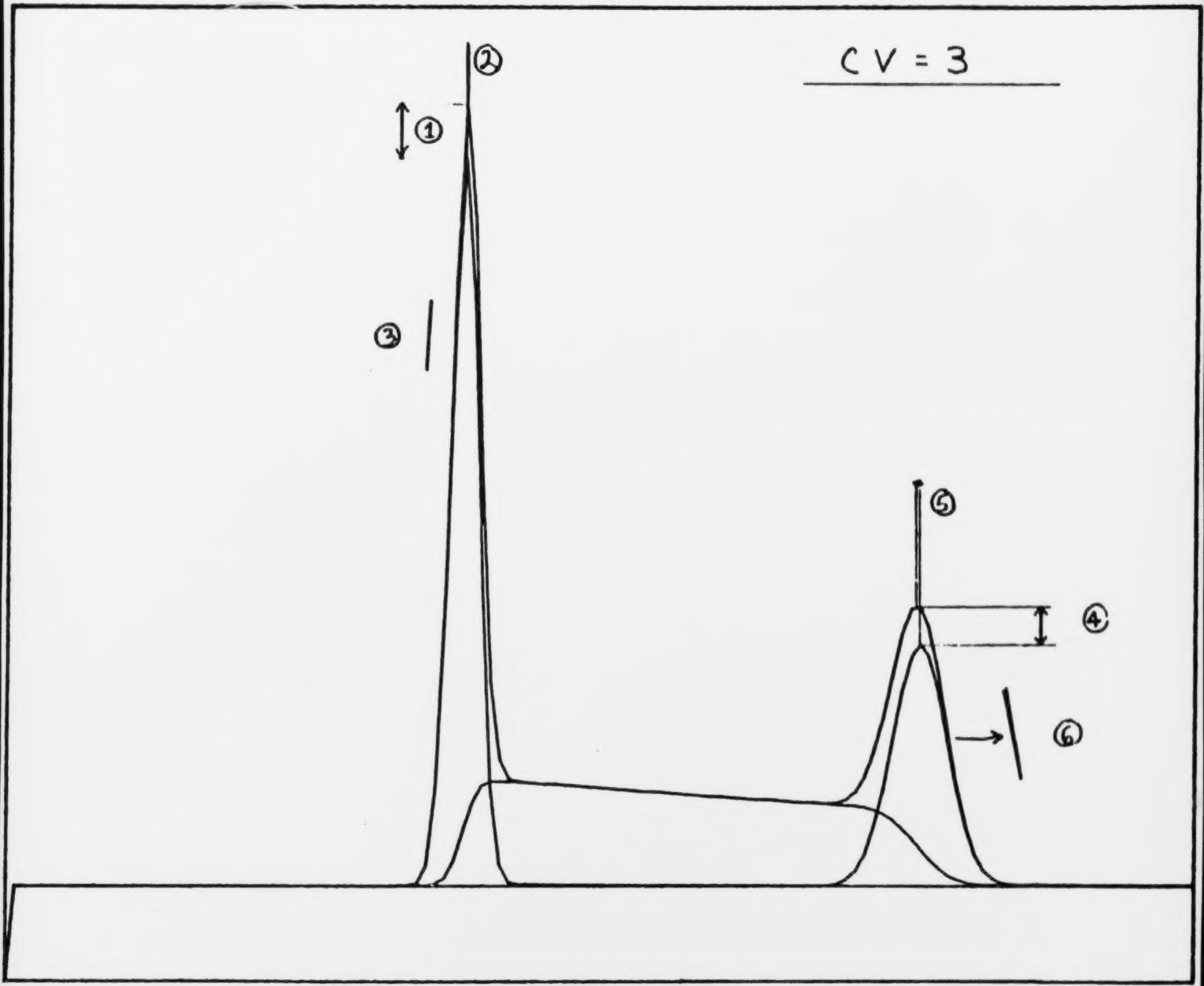

CV = 3
②
①
③
⑤
④
⑥

Figure 6.5-3: Family of error correction curves for the peak height. CV=6, 7, and 8%.

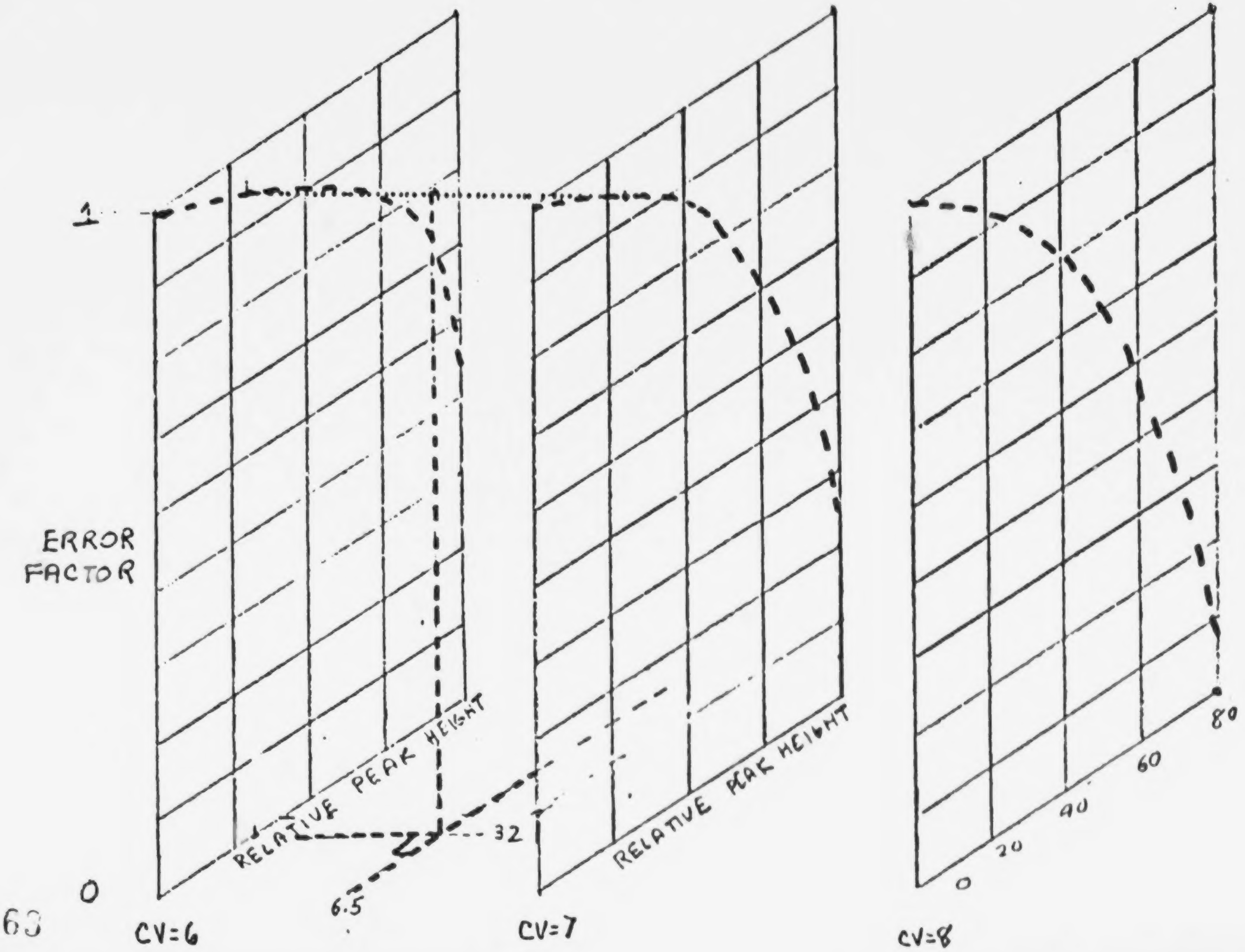
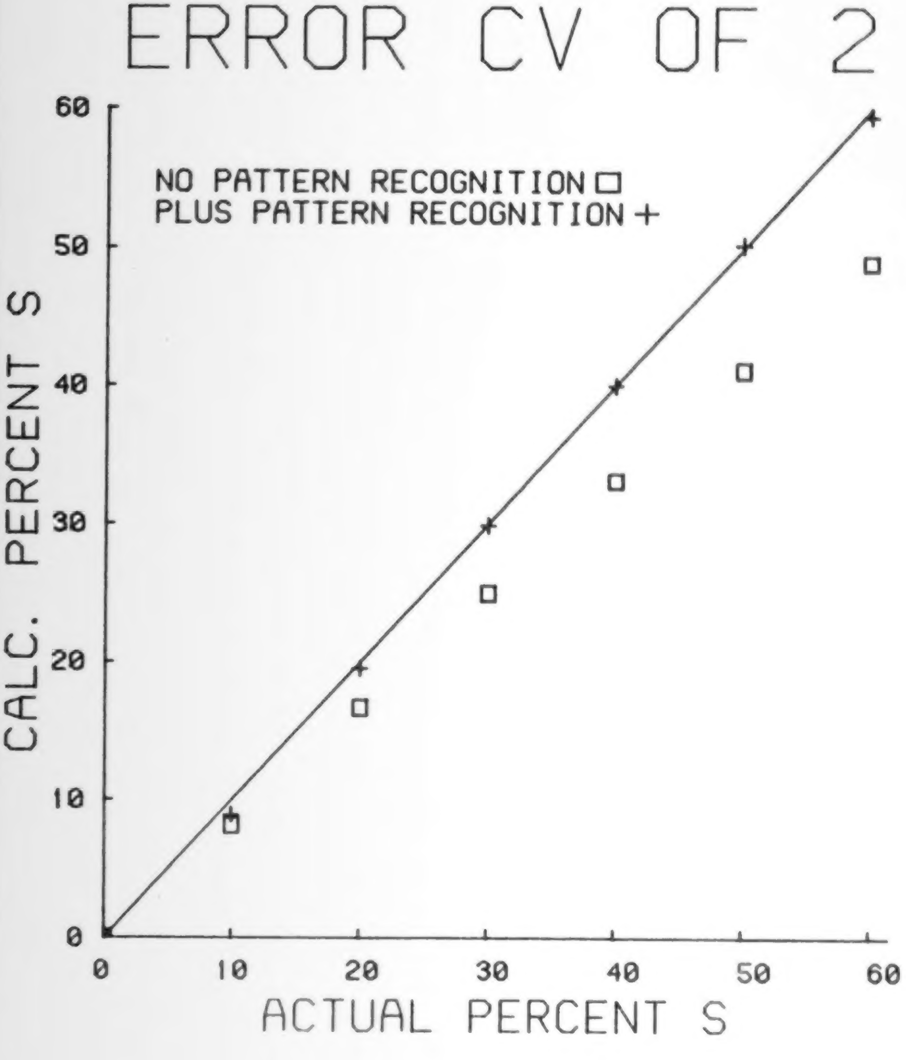
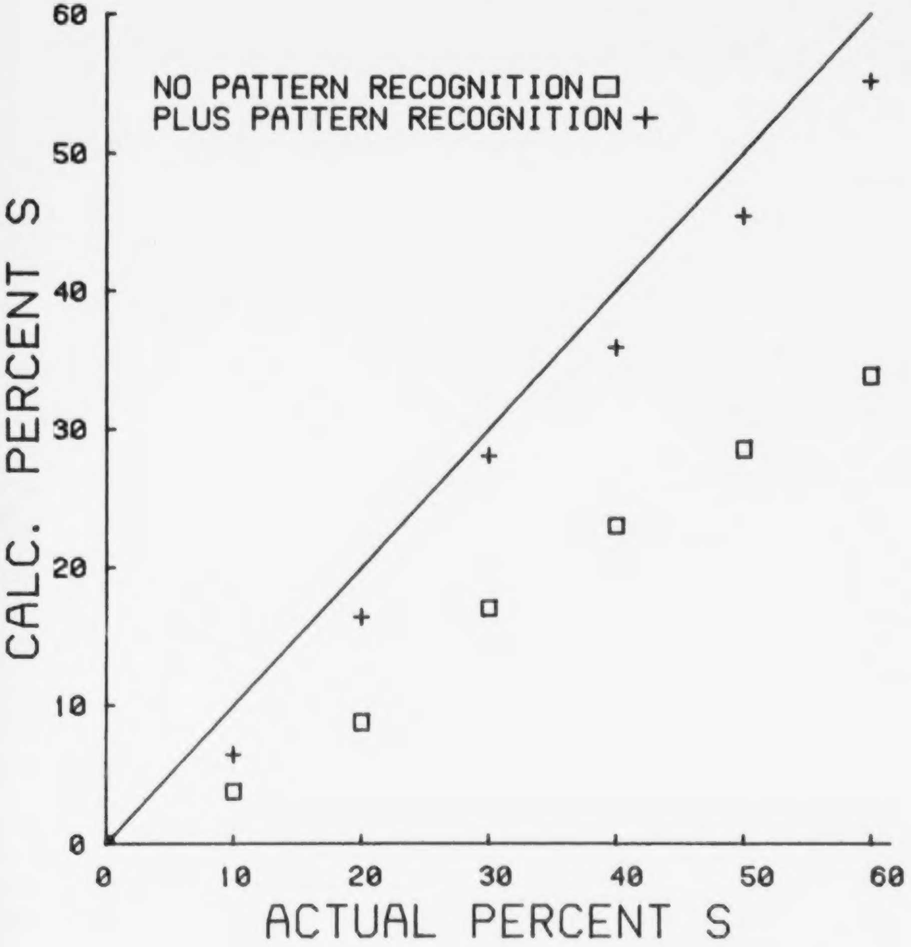

1
ERROR
FACTOR
0
RELATIVE PEAK HEIGHT
32
6.5
CV=6
RELATIVE PEAK HEIGHT
CV=7
0
20
40
60
80
CV=8

Figure 6.5-4: Effect of pattern recognition on estimation of S for A) CV=2% and B) CV=10%.


ERROR CV OF 2
NO PATTERN RECOGNITION □
PLUS PATTERN RECOGNITION +
CALC. PERCENT S
ACTUAL PERCENT S
0
10
20
30
40
50
60
0
10
20
30
40
50
60

ERROR CV OF 10



6.6 Translocation.

6.61 Introduction.

After the histogram has been smoothed and capped, it can be translocated to any desired position. If the user wishes to eventually compare groups of histograms, this step is usually necessary. The histogram position can drift slightly for a number of reasons e.g. variations in staining, temperature, fluid pressure, orifice size, laser output, laser alignment, etc. The easiest way to compensate for these sources of variation is to run an internal standard or key in on some inherent marker. One can use internal standards such as microspheres or avian RBC's. If one is dealing with a DNA histogram, the G0-G1 peak can be used as an inherent marker. This routine allows the user to translocate the histograms such that all the standards or markers line up precisely at any given channel number. By using this technique, the sensitivity of comparing two groups of histograms (nonparametric analysis) is greatly increased.

6.62 Mathematics.

The mathematics associated with this routine are trivial.

The parent histogram is simply mapped in a conventional way to a new axis. Care is taken to maintain a constant area under the curve. The mapping is made possible by interpolating between points using the smoothing routine's parabolic arc coefficients. Therefore, in order to translocate, one must first smooth the data.

6.7 Doublet Correction.

6.71 Introduction.

One of the limitations of flow cytometry is the necessity of making a single cell suspension. If there are aggregates of cells present they will be registered by the flow cytometer as one cell with a large amount of DNA. It is theoretically possible to find a light scatter angle that will differentiate multiple cells cleanly from single cells. Also, time of flight may be valuable in this respect (20).

Another possible approach to limiting the degree of aggregation is in modifying the cell preparation. A large number of variations of the staining procedure have been tried.

1. Use of proteases and nucleases
2. low concentration of DNAse
3. EDTA
4. Ca++
5. Collagenase
6. Ph
7. Salt Concentration
8. Temperature
9. Sonication and vortexing

Some of these modifications to our staining procedure did reduce the the number of aggregates, but none totally eliminated the aggregates for, say, parathyroid tissue. There are some circumstances where even a 1-2% level of doublets is intolerable (e.g. differentiating between normal parathyroids and hyperplastic parathyroids).

I finally decided to try to correct for doublet contamination with a mathematical technique. We have observed that when cells aggregate there are mostly singlets, a smaller number of doublets and a still smaller number of triplets, quadruplets, etc. These aggregates can be observed in the fluorescent microscope or the flow cytometer. If we take a resting population of cells (e.g. parathyroids) with no S or G2-M, and run them through the flow cytometer we obtain a histogram such as the one shown in Figure 6.7-1. In order to appreciate the small peaks representing doublets, triplets and quadruplets, we plot the ordinate in the log domain. This has the effect of minimizing large peaks (e.g. G0-G1) and maximizing small peaks (eg. triplets; see Figure 6.7-2).

After having observed a large number of these histograms, I noticed that there seemed to be some pattern in the numbers of singlets, doublets, triplets and quadruplets. I reasoned that if one were able to linearize the relationship between the number of each aggregate, then one could measure the number of singlets and triplets and interpolate to find the number of doublets. We would then have some way of distinguishing between doublets and G2-M cells. This procedure takes advantage of the fact that normally there are not 6C cells in a cell population and therefore the peak in the 6C region is mostly triplets.

It turns out that for computation reasons, it was best to linearize the peak heights of the singlets, doublets and triplets. This was done by using ten resting parathyroid cell histograms and determining the peak heights of the singlets, doublets, and triplets. This data was entered into a linear regression program. This program did 576 transformations on both the independent (channel number) and dependent (peak height) variables to find transforms that best linearized the data. After performing this analysis on each of the ten different parathyroid histograms, two sets of transforms were found that best linearized the data:

$LN(Y)$ vs $LN(X)/X$ and, $Y^{1/4}$ vs $1/X^2$

Figures 6.7-3 and 6.7-4 show typical results of these two transforms. In order to appreciate the effect of these two transforms on an entire histogram, two programs were written that graphically demonstrate these transforms (see Figures 6.7-5 and 6.7-6). Notice, that I use 1-LN(X)/X instead of

LN(X)/X in order to have the G0-G1 on the left and the aggregates on the right.

I finally decided on the natural log transformation because the peaks were more evenly spaced than the exponent transformation. Also, the parameter of CV cancels out with this transformation thus making it CV independent. This transformation seemed to work very well for parathyroids, but would it work also for other cell types? To answer this question we looked at resting lymph node and ventral prostate nuclei. Figure 6.7-7 shows the transformed histogram of lymph node cells and the linear relationship between singlets, doublet, and triplet peak heights is preserved. Figure 6.7-8 shows the transformed histogram of ventral prostatic nuclei. It was very encouraging to observe that this tissue was highly aggregated, but the linear relationships were still maintained.

In summary, this routine seems to be able to accurately estimate the number of doublets by transforming the histogram in a unique way and interpolating between the singlet and triplet peak heights to find the doublet peak height. One can easily calculate the SD of the peak by either averaging the triplet SD and the singlet SD or by directly measuring the 4C peak SD. Once one has the peak height, peak position, and standard deviation, one can subtract the doublet Gaussian dictated by these parameters from the parent curve and theoretically have a better approximation of the G2-M compartment.

6.72 Mathematics.

Let Ps, Pd, and Pt be the peak positions of the singlet, doublet, and triplet peaks respectively. Also let Hs, Hd, and Ht be the peak heights of the singlet, doublet and triplet peaks respectively. Figure 6.7-9 schematizes the geometric relationships between these variables. To determine the doublet height we take advantage of the fact that triangle ABC is congruent with triangle ABD. Therefore the ratios of the sides of the two triangles are equal.

$$\frac{LN(Hs) - LN(Ht)}{\frac{LN(Ps)}{Ps} - \frac{LN(Pt)}{Pt}} = \frac{LN(Hs) - LN(Hd)}{\frac{LN(Ps)}{Ps} - \frac{LN(Pd)}{Pd}}$$

Solving for LN(Hd)

$$LN(Hd) = LN(Hs) - \frac{(LN(Hs)-LN(Ht))*\left(\frac{LN(Ps)}{Ps}-\frac{LN(Pd)}{Pd}\right)}{\frac{LN(Ps)}{Ps} - \frac{LN(Pt)}{Pt}}$$

Figure 6.7-1: DNA histogram of resting parathyroid cells.

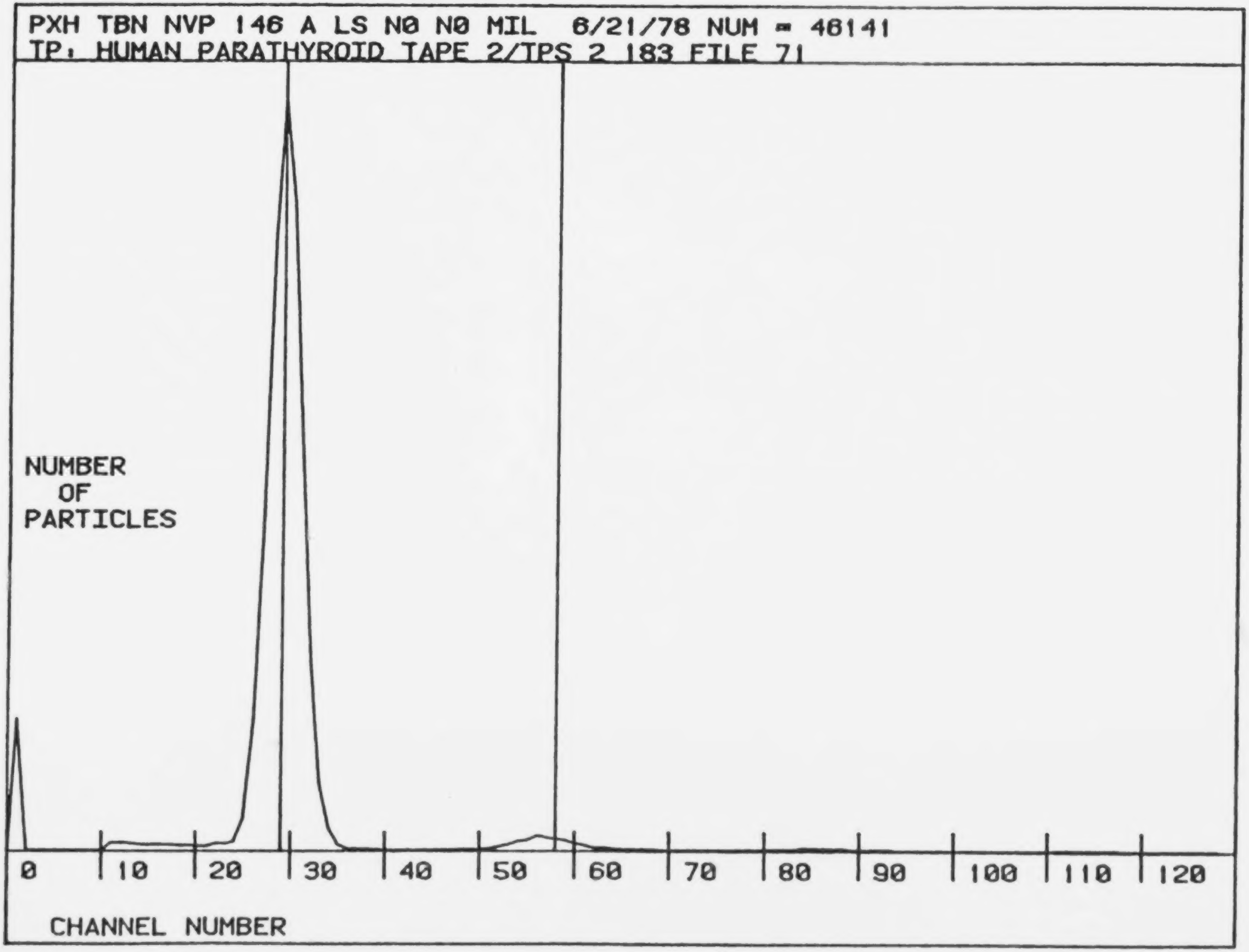
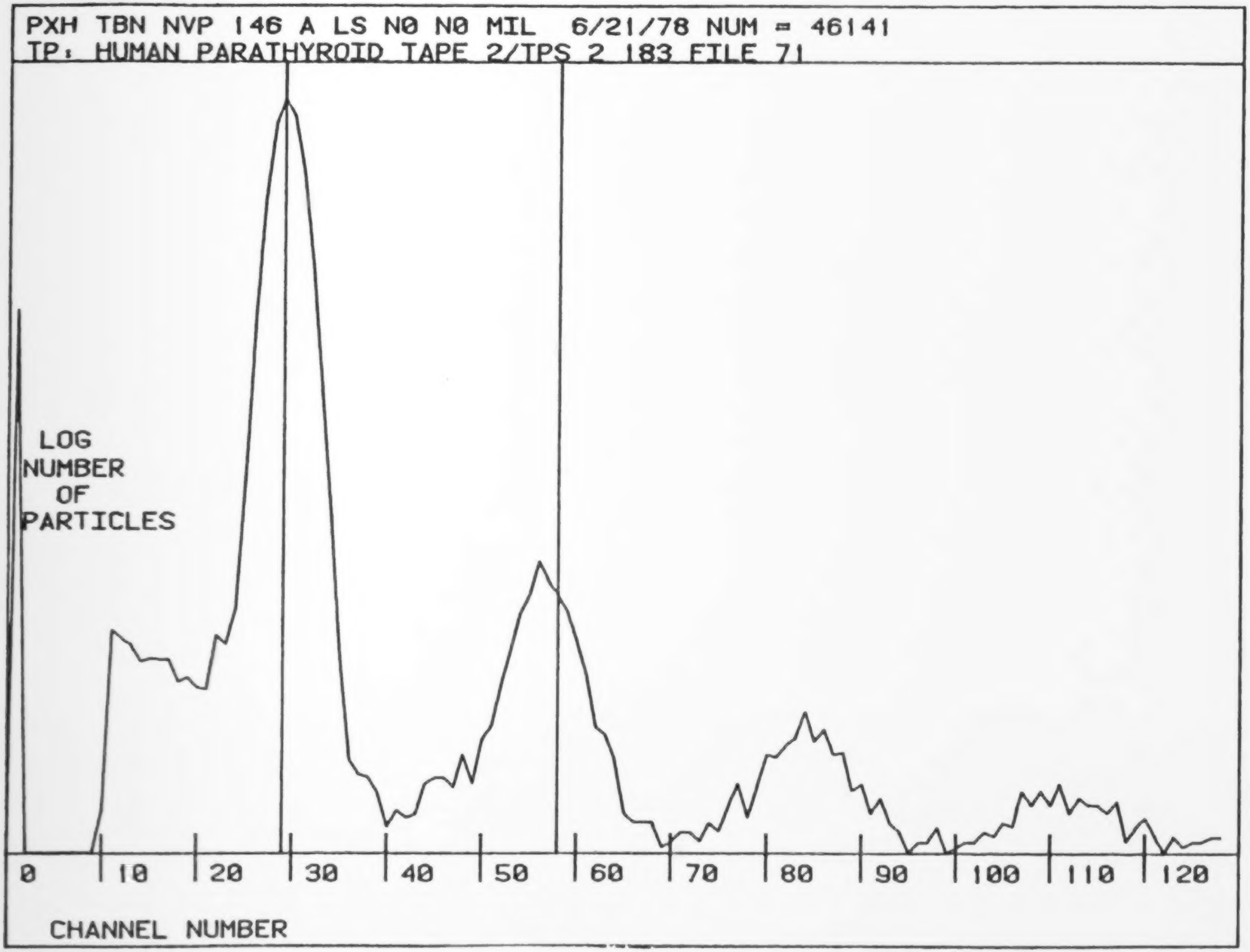

PXH TBN NVP 146 A LS N0 N0 MIL 6/21/78 NUM = 46141
TP: HUMAN PARATHYROID TAPE 2/TPS 2 183 FILE 71
NUMBER OF PARTICLES
0
10
20
30
40
50
60
70
80
90
100
110
120
CHANNEL NUMBER

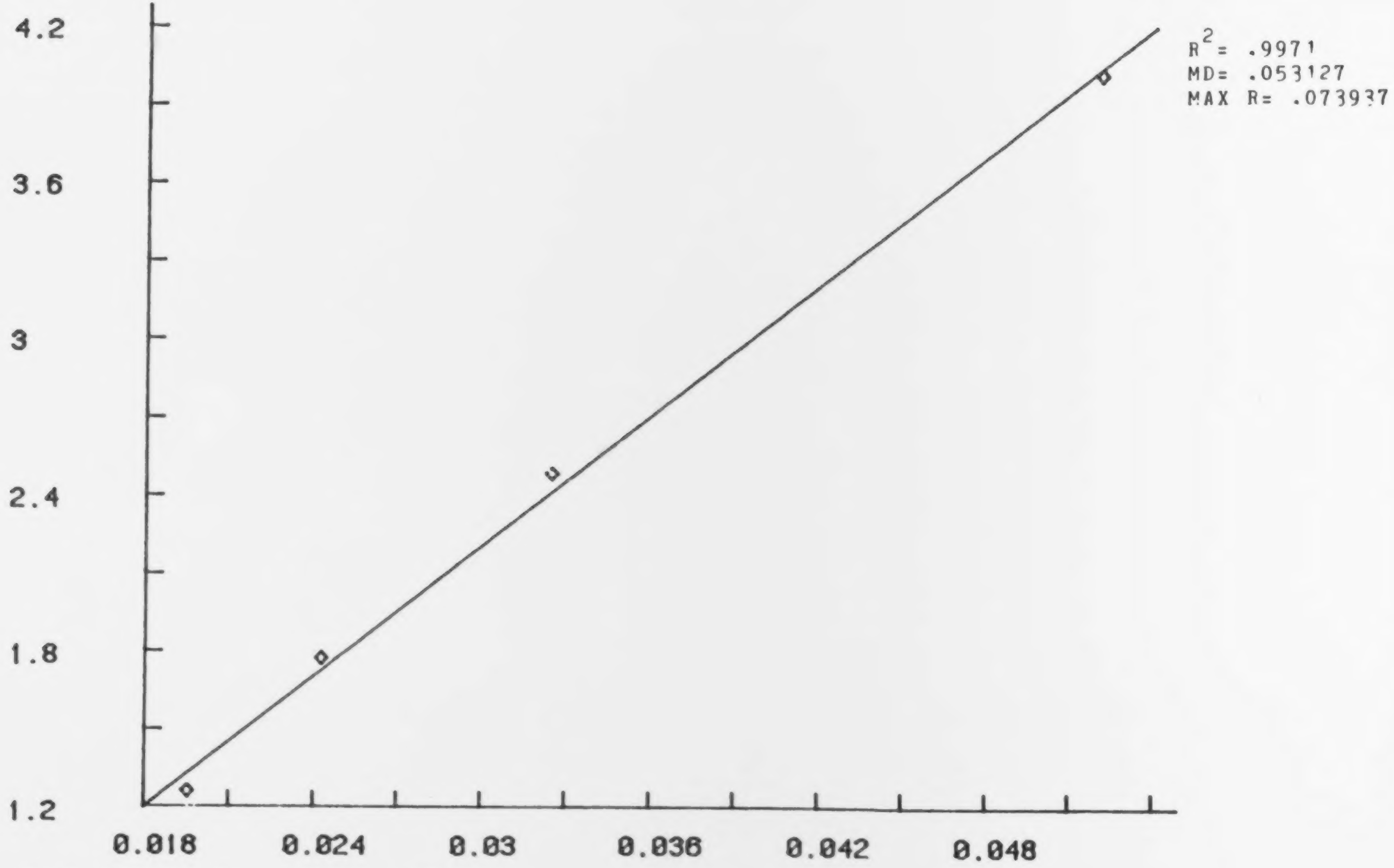
Figure 6.7-2: Semilog of Figure 6.7-1.


PXH TBN NVP 146 A LS N0 N0 MIL 6/21/78 NUM = 46141
TP: HUMAN PARATHYROID TAPE 2/TPS 2 183 FILE 71
LOG NUMBER OF PARTICLES
0
10
20
30
40
50
60
70
80
90
100
110
120
CHANNEL NUMBER

170

180

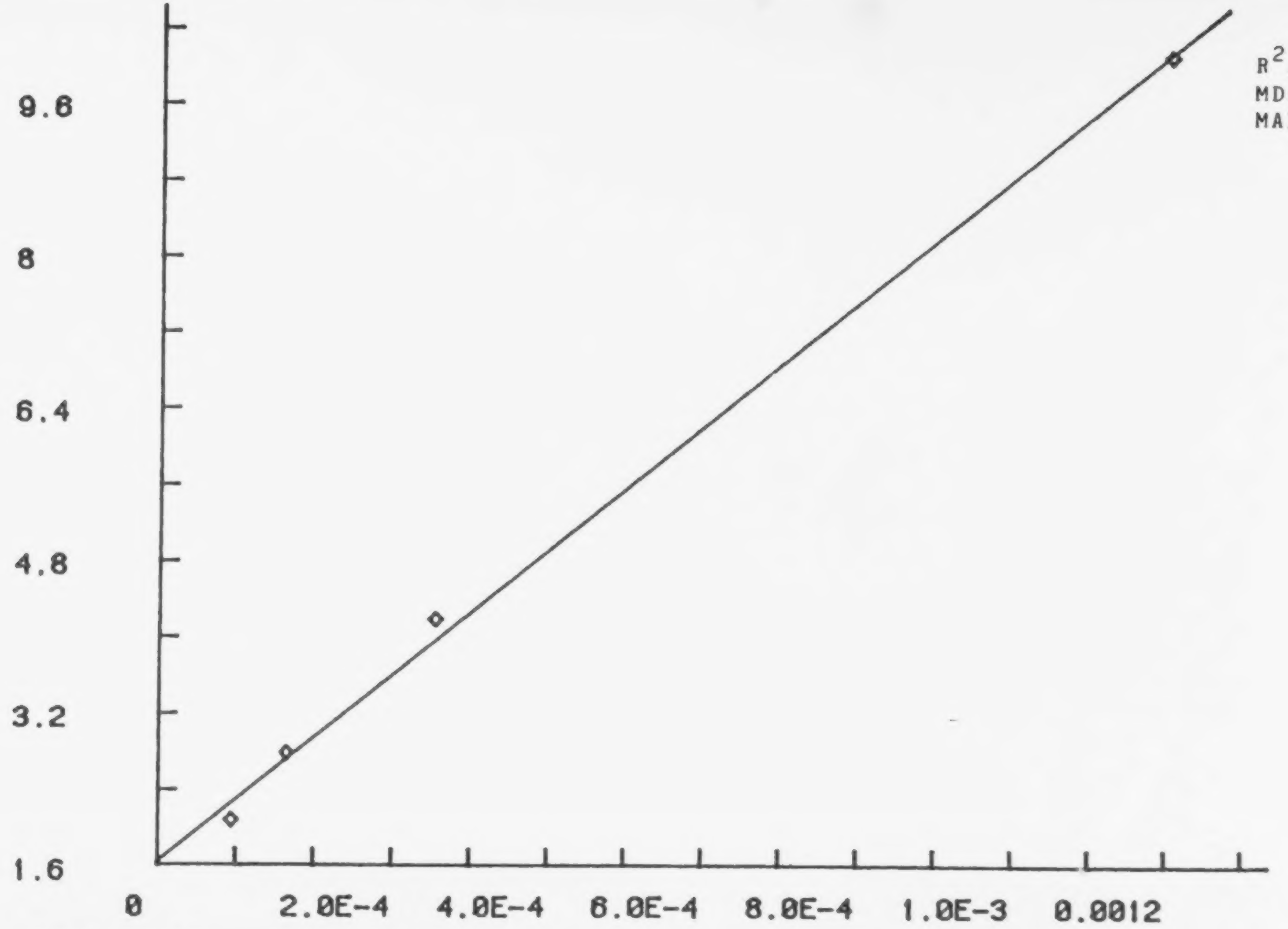
Figure 6.7-3: Log aggregation transform of parathyroid singlet, doublet, and triplet peak heights.



F(X)=LGT(V)/V   G(Y)=LGT(V)

$G(Y) = -0.30682 + 83.64389 * F(X)$

Figure 6.7-4: Power aggregation transform of parathyroid singlet, doublet, and triplet peak heights.



R^2= .9978
MD= .123601
MAX R= .205672

F(X)=1/V^2   G(Y)=V^(1/4)

G(Y) =   1.64386 + 6521.27849 * F(X)

Figure 6.7-5: Log aggregation transform of entire parathyroid histogram.

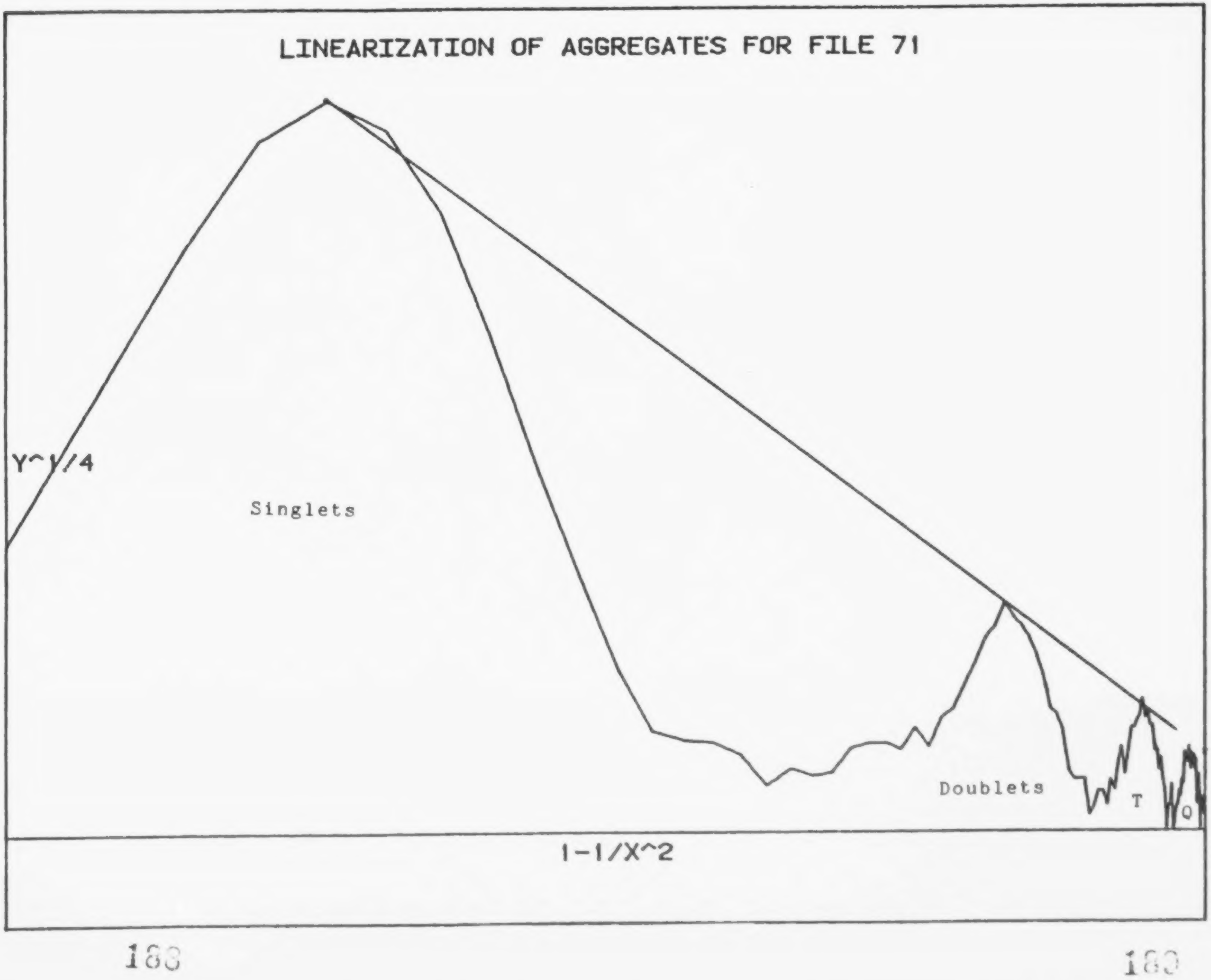

LINEARIZATION OF AGGREGATES FOR FILE 71
Y^1/4
Singlets
Doublets
T
Q
1-1/X^2
188
189

Figure 6.7-6: Power aggregation transform of entire parathyroid histogram.

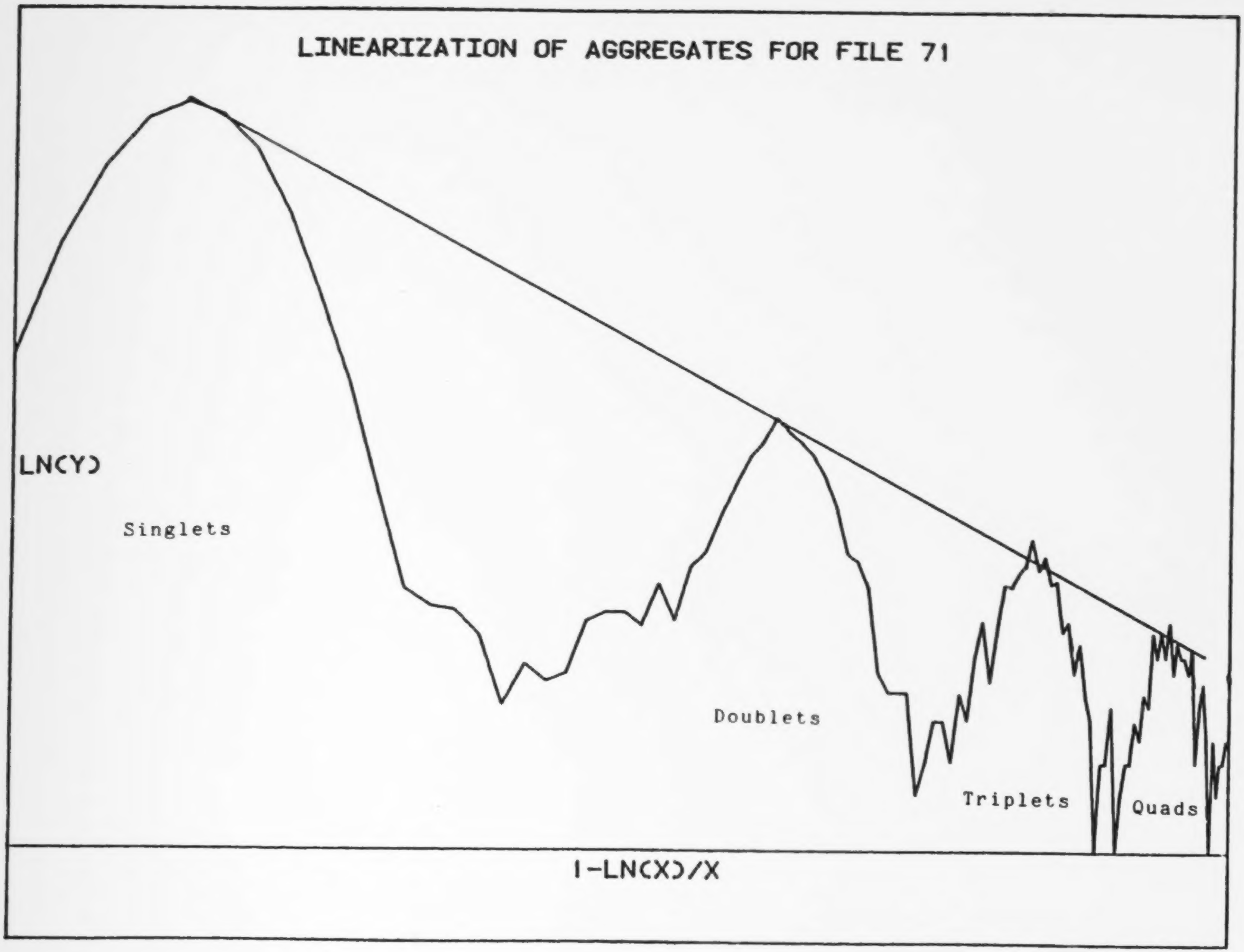
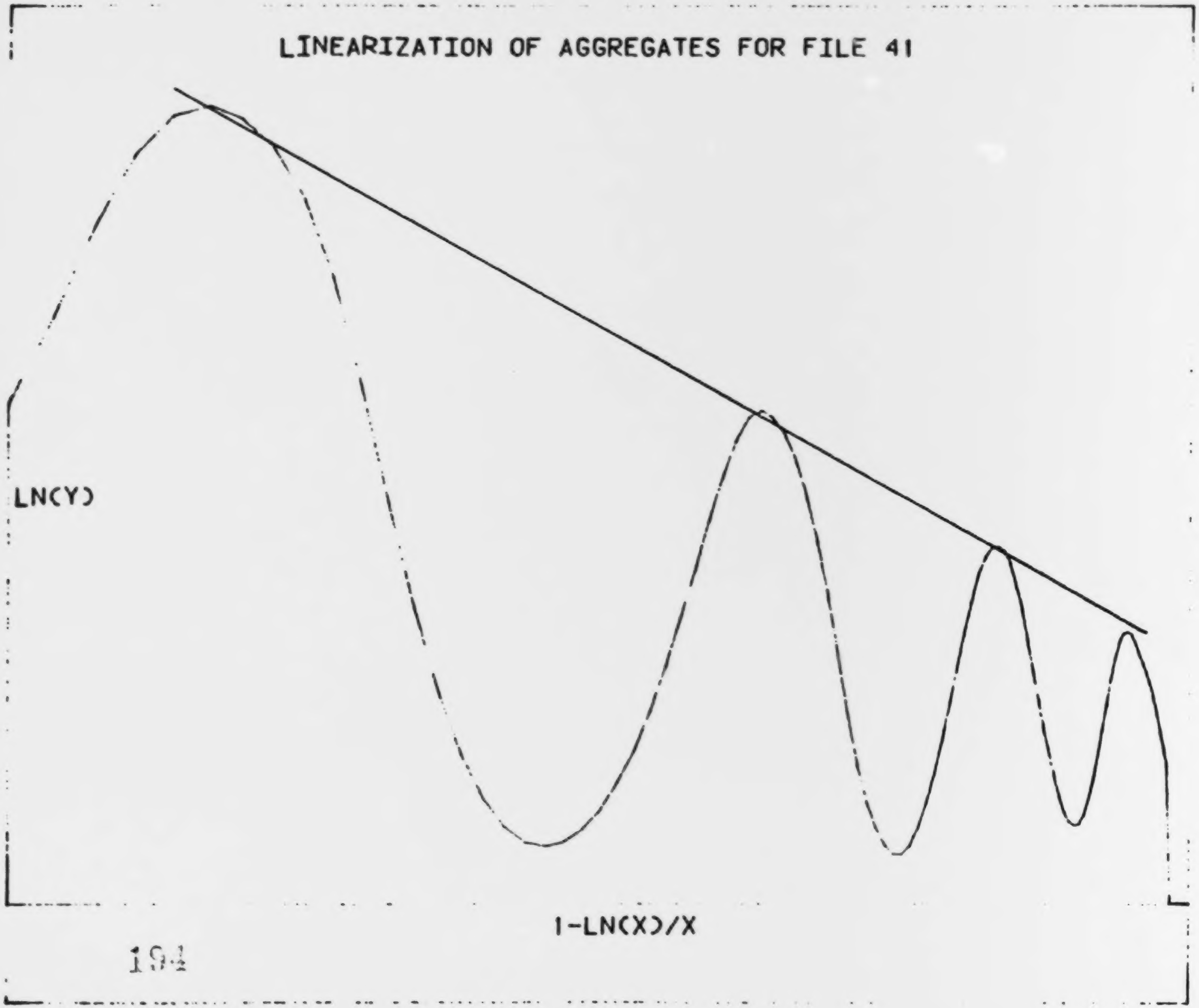

LINEARIZATION OF AGGREGATES FOR FILE 71
LN(Y)
Singlets
Doublets
Triplets
Quads
1-LN(X)/X

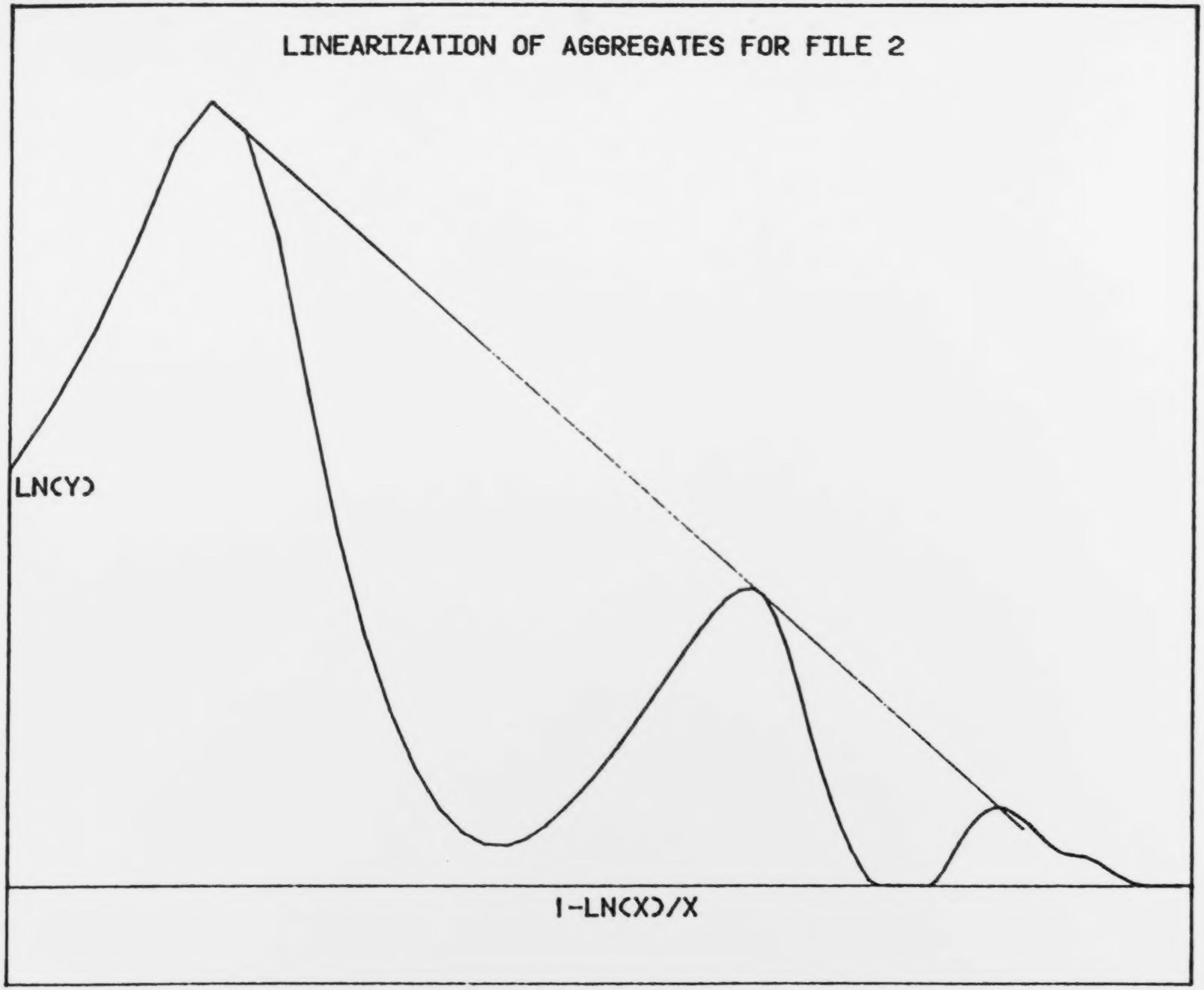
Figure 6.7-7: Log aggregation transform of lymph node cells.

Prostate.



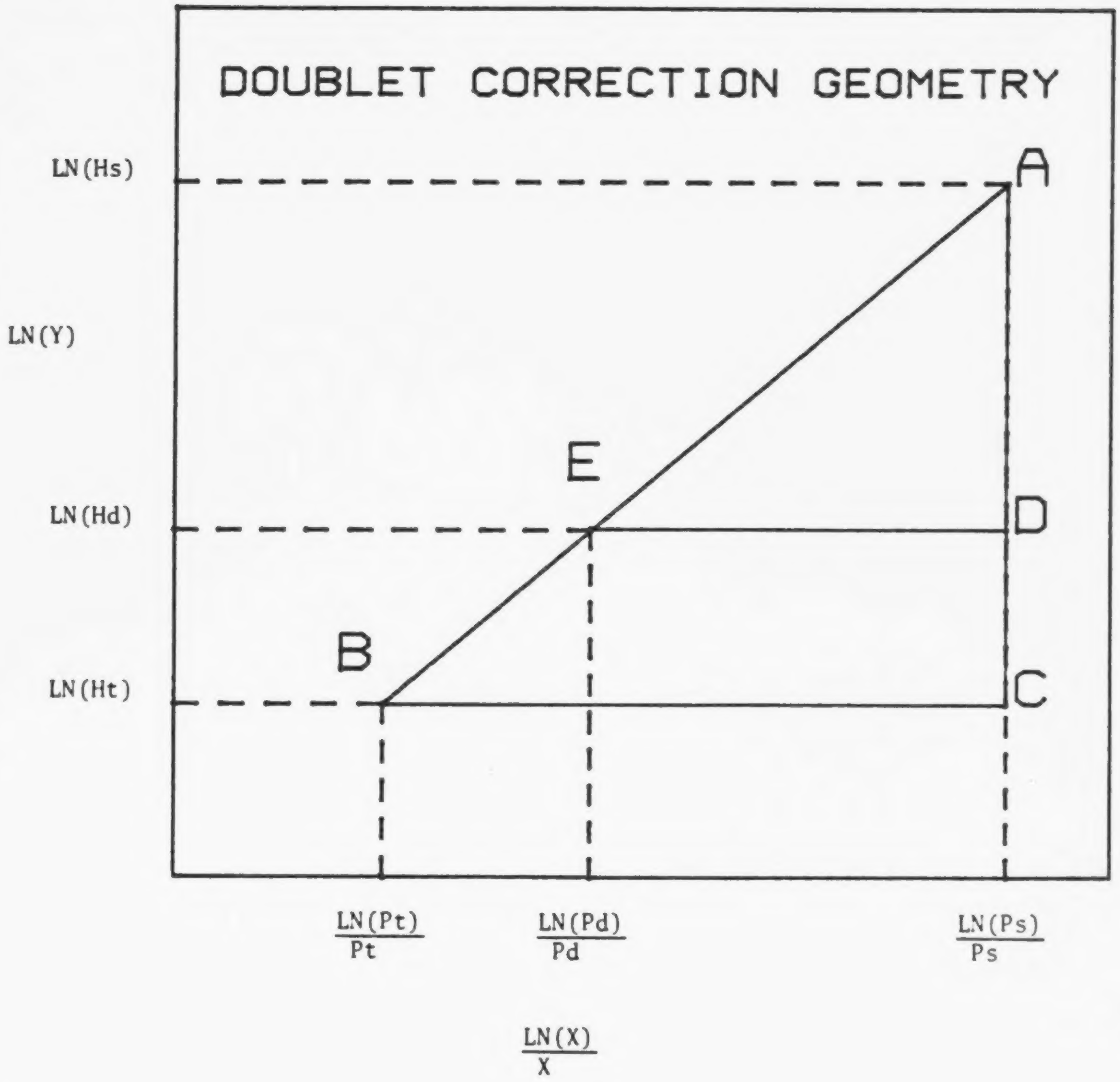
3. 6

Figure 6.7-8: Log aggregation transform of ventral prostate cells.



LINEARIZATION OF AGGREGATES FOR FILE 2

Figure 6.7-9: Geometric relationships between transformed variables.


DOUBLET CORRECTION GEOMETRY
LN(Hs)
A
LN(Y)
E
LN(Hd)
D
B
LN(Ht)
C
$\frac{LN(Pt)}{Pt}$
$\frac{LN(Pd)}{Pd}$
$\frac{LN(Ps)}{Ps}$
$\frac{LN(X)}{X}$

7. Parametric analysis II.

7.1 Introduction.

The output from the crude stripping and pattern recognition adjustment routines can be used as estimates for parametric analysis II. Parametric analysis II uses a Marquardt nonlinear least squares algorithm (Chap 7.2) to obtain the most probable kernel (Chapts 7.3-7.6) that describes the data. The kernel or model can then give directly the biological parameters of percent G0-G1, S, and G2-M or indirectly the parameters of relative transit times and DNA synthesis rates (Chapt 7.622). The next chapter will discuss some of the theoretical aspects of nonlinear least squares, especially the one proposed by Marquardt in 1963.

7.2 Nonlinear least squares analysis (21).

7.21 Introduction.

In Chapters 6.3 and 6.4 we used linear least squares analysis to find the best Gaussian through a number of points. The process was linear because it could be transformed to an equation such as

$$Y = P0 + P1X1 + P2X2 + \ldots + PmXm + e \quad (7.2\text{-}1)$$

where Pi are the parameters, Xi are the independent variables, and e is the error. A word of caution should be made about transformations before continuing. Let us say we have a model such as

$$Y = e^{-KX^2} \quad (7.2\text{-}2)$$

that we wish to fit to some data. By taking the log of both sides, the equation transforms to

$$LN(Y) = -KX^2$$

If we let $LN(Y)=Y'$, $X^2=X1$, and $-K=P1$ we then have

$$Y' = P1X1$$

This equation, of course, is linear and we can thus subject it to a linear least squares analysis. This analysis minimizes the SSQ between the model and the data in log space to determine the best P1. This procedure yields the least SSQ in log space, but not necessarily in linear space.

In fact, in this example the calculated P1 value will be underestimated by the least squares analysis in log space. The degree of underestimation will be a function of the spread of points above and below the calculated line. The more dispersion, the greater the error. Therefore, the transformation of intrinsically linear models is not an innocuous procedure and will result in some slight errors.

Any model that cannot be transformed to the form of equation 7.2-1, is called a nonlinear model, that is nonlinear in the parameters. The models that are currently being used in DNA histogram analysis are of this nonlinear form. Let us refer to the complicated nonlinear models as some function (F) with a number of parameters (P) of an independent variable (X). Therefore, the observed data points are approximated by

$$Y = F(P;X) + e \quad (7.2\text{-}3)$$

where e is the error.

In least squares analysis we wish to find the set of parameters (P) that give the least sum of the squares. The equation that calculates the sum of the squares is

$$SSQ = \sum_{k=1}^{n} (Yk - F(P;X))^2$$

By differentiating this equation with respect to each parameter and setting the results equal to zero, we obtain an equation for each parameter.

$$\sum_{k=1}^{n} (Yk - F(P;X)) * \frac{dF(P;X)}{dp1} = 0$$

$$\sum_{k=1}^{n} (Yk - F(P;X)) * \frac{dF(P;X)}{dp2} = 0$$

$$\vdots \qquad\qquad\qquad\qquad\qquad \vdots$$

$$\sum_{k=1}^{n} (Yk - F(P;X)) * \frac{dF(P;X)}{dpm} = 0$$

where m is the number of parameters and dF is the partial derivative.

If F(P;X) were a linear function such as equation 7.2-1, this set of equations could be put in matrix form and solved for the set of parameters (P). Since F(P;X) is nonlinear, we must resort to a number of other methods to find the "best" P that yields the least SSQ. Three such methods are (1) linearization, (2) steepest descent, and (3) Marquardt's compromise. All three of these techniques involve an iterative technique where the computer changes the parameters in some rational manner to decrease the SSQ. The computer will stop iterating when (1) the change in the SSQ is less than some predefined value; (2) the change in the parameters is less than a predefined value; or (3) the algorithm fails for some reason.

7.22 Linearization.

If we could approximate F(P;X) with a linear function, we could then apply the technique of linear least squares analysis to yield a new set of parameters. The approximation that is generally used is the first two terms of the Taylor expansion series. For one parameter the equation is given by

$$Y = F(P0;X) + \frac{dF(P;X)}{dP} * (P1-P0) \qquad (7.2\text{-}3)$$

where P0 is the initial estimate and P1 is a closer estimate to the actual parameter P. Let us see how this approximation works for a simple case, say

$$Y = e^{P*X^{2}} + P \qquad (7.2\text{-}4)$$

Suppose we observe that Y=100 when X=5 and we wish to calculate P using the Taylor expansion series approximation. We first rewrite equation 7.2-3 into a more usuable form

$$P1 = \frac{Y - F(P0;X)}{F'(P0;X)} + P0 \qquad (7.2\text{-}5)$$

where F' is the derivative of F. For this case P1 is calculated as

$$P1 = \frac{100 - e^{P0*25} - P0}{e^{P0*25} * 25} + P0$$

If we let P0 be, say, .5 to start off, P1 would equal .460. In fact, if we then let P1 be the initial estimate and repeat this process, we will converge on the final solution of P with arbitrary accuracy (see Table 7.2-1).

Table 7.2-1: Linearization technique for solving nonlinear problems.

Iteration num	Estimate	Percent error
1	.4600	149.8
2	.4200	128.1
3	.3802	106.5
4	.3405	84.9
5	.3013	63.6
6	.2634	43.0
7	.2289	24.3
8	.2020	9.7
9	.1876	1.9
10	.1843	.1
11	.1841	.0

Equation 7.2-3 can be rewritten to incorporate this iterative process.

$$Y = F(Pj;X) + \frac{dF(Pj;X)}{dP} * (Pj+1 - Pj) \qquad (7.2\text{-}6)$$

for jth iteration.

The Tailor expansion series can be expanded about more than one parameter. The general equation for the first two terms is

$$Y = F(Pj1,Pj2,\ldots,Pjm;X) + \sum_{i=1}^{m} \frac{d(F(P;X))}{dPi} * (Pj+1,i - Pji) \qquad (7.2\text{-}7)$$

We can then substitute this equation into the least squares formulation to yield

$$SSQ = \sum_{k=1}^{n} \left| Y_k - F(P_j;X_k) - \sum_{i=1}^{m} \frac{dF(P_{ji};X_k)}{dP_i} * (P_{j+1,i} - P_{ji}) \right|^2 \qquad (7.2\text{-}8)$$

for the jth iteration.

If we let

$$F_k = F(P_j;X_k)$$
$$B_i = P_{j+1,i} - P_{ji}$$
$$Z_{ik} = \frac{dF(P_{ji};X_k)}{dP_i}$$

then equation 7.2-8 can be rewritten as

$$SSQ = \sum_{k=1}^{n} \left| Y_k - F_k - \sum_{i=1}^{m} B_i * Z_{ik} \right|^2 \qquad (7.2\text{-}9)$$

This equation is now in the form where we can apply the method of linear least squares analysis. In order to best appreciate how this equation transforms to a set of soluble set of simultaneous equations, let us set the number of parameters equal to 3. Therefore, we have

$$SSQ = \sum_{k=1}^{n} \left| Y_k - F_k - B_1 * Z_{1k} - B_2 * Z_{2k} - B_3 * Z_{3k} \right|^2 \qquad (7.2\text{-}10$$

If we take the partial derivative of this equation with respect to each parameter (B), we end up with three equations

$$\sum_{k=1}^{n} (Yk - Fk)*Z1k = \sum_{k=1}^{n} Z1k*Z1k + \sum_{k=1}^{n} Z2k*Z1k + \sum_{k=1}^{n} Z3k*Z1k$$

$$\sum_{k=1}^{n} (Yk - Fk)*Z2k = \sum_{k=1}^{n} Z1k*Z2k + \sum_{k=1}^{n} Z2k*Z2k + \sum_{k=1}^{n} Z3k*Z2k$$

$$\sum_{k=1}^{n} (Yk - Fk)*Z3k = \sum_{k=1}^{n} Z1k*Z3k + \sum_{k=1}^{n} Z2k*Z3k + \sum_{k=1}^{n} Z3k*Z3k$$

To make these equations simpler to work with, let us adopt the following notation.

$$[Y1] = \sum_{k=1}^{n} (\underline{Y}k - Fk)*Z\underline{1}k$$

$$[12] = \sum_{k=1}^{n} (Z\underline{1}k * Z\underline{2}k)$$

etc.

The above system of equations neatly transforms to

$$[Y1] = B1*[11] + B2*[21] + B3*[31]$$
$$[Y2] = B1*[12] + B2*[22] + B3*[32]$$
$$[Y3] = B1*[13] + B2*[23] + B3*[33]$$

All the numbers in brackets can be determined; therefore, we have three equations and three unknowns. The solution of these simultaneous equations (B1,B2,B3) can be calculated by multiplying the inverted coefficient matrix by the resultant vector, that is

$$\begin{bmatrix} {[B1]} \\ {[B2]} \\ {[B3]} \end{bmatrix} = \begin{bmatrix} {[11]} & {[21]} & {[31]} \\ {[12]} & {[22]} & {[32]} \\ {[13]} & {[23]} & {[33]} \end{bmatrix}^{-1} * \begin{bmatrix} {[Y1]} \\ {[Y2]} \\ {[Y3]} \end{bmatrix}$$

The solution set, B (B1,B2,B3), can be used to calculate (Pj+1,1) ,(Pj+1,2),and (Pj+1,3). This process can be repeated until the SSQ does not change appreciably. The final set of parameters is then arbitrarily close to the true parameters: P1, P2, and P3. One can easily see how this process can be extended to m parameters.

It often happens, as in our case, that the function we are fitting is so complex that the closed form derivative is unknown. We can approximate the derivative by changing the parameter by some small amount and observing the change in the dependent variable. If the change in the parameter is small, this is a good approximation. The linearization technique does have some possible drawbacks.

1. It may converge very slowly; that is, a very large number of iterations may be required before the solution stabilizes even though the SSC may decrease as j (iteration number) increases.

2. It may oscillate widely, continually reversing direction, and often increasing, as well as decreasing, the SSC. Nevertheless, the solution may stabilize eventually.

3. It may not converge at all so that the SSC increases iteration after iteration without bound.

The next chapter will discuss a method that uses an entirely different method of finding parameters yielding the minimum SSC.

7.23 Steepest descent (21).

The steepest gradient or descent method has a beautifully simple geometric interpretation. If you imagine the SSQ's function as some hilly terrain, what we are after is the lowest valley which would, of course, correspond to the lowest SSQ's. Like water flowing downhill, this method follows the contour of this terrain to its lowest point. The computer simply evaluates the slope in each parameter direction and heads towards the steepest incline with some artitrary step size. If the slope is negative, we are heading in the right direction and the parameter is increased some positive amount which will be a function of some scale factor. Suppose we have a function with three parameters. The equations that would iteratively converge to the SSQ's valley would be

$$P_{j+1,1} = P_{j,1} + S1 * -\left(\frac{dSSQ}{dP1} \right)$$

$$P_{j+1,2} = P_{j,2} + S2 * -\left(\frac{dSSQ}{dP2} \right)$$

$$P_{j+1,3} = P_{j,3} + S3 * -\left(\frac{dSSQ}{dP3} \right)$$

where S1, S2, and S3 are artitrary scale factors. There are a number of analytical expressions on how to calculate S1, S2, and S3 for optimal convergence; however, we wish to use the scales that Marquardt derived in order to have a smooth transition to the next chapter. If we use the same notation conventions that were adopted in the preceding chapter, these equations become

$$[Y1] = B1 * \frac{-2}{S1}$$

$$[Y2] = B2 * \frac{-2}{S2}$$

$$[Y3] = B3 * \frac{-2}{S3}$$

You will see in a moment the advantage in letting S1 be calculated as

$$S1 = \frac{-2}{[11]*(1+L)}$$

where L is lambda and its significance will become apparent in the next chapter. If we make similar definitions for S2 and S3 we obtain the following equations:

$$\begin{aligned}
[Y1] &= B1 * [11] * (1+L)\\
[Y2] &= B2 * [22] * (1+L)\\
[Y3] &= B3 * [33] * (1+L)
\end{aligned}$$

These equations should be compared to the equations 7.2-12 in the preceding chapter. The advantage of this technique is that it will always converge; however, in the relatively flat region near the lowest SSC's, it convergence slows down considerably. It was Marquardt who realized that the techniques of linearization and steepest gradient could be combined to maximize their separate advantages and minimize their disadvantages.

7.24 Marquardt compromise (52).

In 1963 Marquardt published a method (22) that combined the best features of both the linearization and steepest gradient nonlinear least squares analysis techniques. The linearization technique has an extremely quick convergence near the lowest SSQ's value whereas the steepest gradient technique has a slow convergence at this point. At a point relatively far from the SSQ's value, the linearization of the function may be a very poor approximation and therefore the technique may not converge; however, the steepest gradient technique will always converge and is particularly good at the periphery of the valley. It is clear that a combination of these two techniques would be a powerful tool in analyzing nonlinear problems.

In chapter 7.22 we developed a system of simultaneous equations for the linearization techique. They were

```
[Y1] = B1*[11] + B2*[21] + B3*[31]
[Y2] = B1*[12] + B2*[22] + B3*[32]
[Y3] = B1*[13] + B2*[23] + B3*[33]
```

and in chapter 7.23 we developed a system of equations for the steepest gradient technique.

```
[Y1] = B1*[11]*(1+L)
[Y2] =           B2*[22]*(1+L)
[Y3] =                          B3*[33]*(1+L)
```

We can rewrite the linearization equations as

```
[Y1] = B1*[11]*(1+L) + B2*[21]        + B3*[31]
[Y2] = B1*[12]       + B2*[22]*(1+L)  + B3*[32]
[Y3] = B1*[13]       + B2*[23]        + B3*[33]*(1+L)
```

For small lambdas (L<<1), this system of equations approaches the linearization technique. For large lambdas (L>>1), the diagonal of this matrix becomes dominant and the equations approach the steepest gradient technique. It should be noted that Marquardt normalizes the equations such that the diagonal is unity before he evaluates the inverse. Thus, by changing lambda one can smoothly change from one technique to another. The simple recipe given by Marquardt for changing L is

1. Compute the SSQ's (SSQi) at the initial estimate.
2. Start with lambda equal to .001 (linearization predominates).
3. Compute the parameters (B) and determine the new SSQ's (SSQ(i+1)).
4. If SSQ(i+1)≥ SSQi then increase lambda by a factor of 10 and repeat step 3.
5. If SSQ(i+1)<SSQi then decrease lambda by a factor of 10, consider the new parameters as estimates and return to step 3.
6. Stop the procedure when the convergence criteria are met.

The procedure is simple yet will rapidly converge to the minimum SSQ's. The validity of the final result, however, depends on the model used. The model for this case must have firm biologic confirmation. All too often, models are judged on their ability to fit the data - the better the fit, the better the model. This kind of reasoning can quickly be disposed of by the simple observation that a 128th degree polynomial will fit a DNA histogram perfectly, but what does it tell us about the biology? Nothing!

There are at least three major areas where assumptions are made in models for DNA histogram analysis: 1) nature of the cycling state of the cells, 2) nature of the spread function, and 3) the nature of fluorescent intensity. The cells will be assumed to be in a relative asynchronous state for our analysis. There are two aspects of the spread function that will be discussed in further detail: is noise normally or lognormally distributed (Chapter 7.3), and what is the relationship between standard deviation and channel number? (Chapter 7.4). The last assumption that will be discussed will involve the relationship between stain intensity and DNA content (Chapter 7.5).

7.3 Normal vs lognormal noise distribution.

In 1976, Pearlman presented theoretical work demonstrating that the spread distribution of a single message in a DNA histogram was lognormally distributed (23). The operational difference between these two assumptions is that in the lognormal case, the intensity axis is transformed to the log domain and then treated in the same manner as Fried (multigaussian nonlinear least squares). Fried has stated (25) that when he has tried both techniques on the same distribution, there appears very little difference in the final parameter values. The only hard data that I've seen on this matter is a comparison performed by Brunsting (26) with fluorescent bead preparations. Ten different unimodal histograms were analyzed by both normal and lognormal models of the data. The results showed no clear difference between these two methods.

Although this assumption is at the very foundation of DNA histogram analysis, it apparently makes little difference which way you go. We, therefore, assume a Gaussian distribution mainly because of its simplicity and ease of use.

7.4 Standard deviation vs channel number.

Up until 1976, most DNA histogram analyses were done assuming that CV was constant. It was apparent to Brunsting and this author that there were a large number of histograms where this assumption did not seem valid. In collaboration with Kane and Collins at Virginia Commonwealth University, we studied (27) two different biologic systems (mouse liver and rat testis) to ascertain what was the best simple functional relationship between standard deviation and channel number. These systems were chosen because they both had three distinct peaks (Liver: 2C, 4C, 8C; Testis: 1C, 2C, 4C) and they both had very little S phase. A total of eight simple models were chosen with at most two free parameters. These eight models were applied to the dependent variables CV and SD; therefore, there were a total of 16 different possible models describing CV or SD vs channel number. The CV's and SD's for eight mouse liver and eleven rat testis were analyzed by a three Gaussian weighted nonlinear least squares analysis. The uncertainties associated with each model were calculated and the models were ranked on the final chi square value. The results were that the linear relationship between SD and channel number was the best overall model, that is

$$SD = MX + B$$

where X is the channel number.

It was interesting to note that in all cases, the assumption that CV is constant with respect to channel number was found to be significantly poorer than most other models. These tests suggested that the DNA histogram analyses which assumed only a linear relation between SD and channel number, were better biologically posed than constant CV models. The constant CV models also imply that there is a perfect 2C-4C relationship between the G0-G1 and G2-M means. In the next chapter we will discuss why this ideal relationship may not be the case.

7.5 DNA/stain content vs fluorescent intensity.

Stain content (eg. PI) is not directly proportional to fluorescence (6). The following discussion will show why the observed G2-M peak is sometimes observed slightly to the left of the calculated position. If a computer program is employed that assumes a perfect 2C-4C relationship, considerable error can be encountered in the estimation of G2-M and late S. This deviation from ideality could be explained by stoichiometric changes in the DNA when cells enters G2-M. This change could also be on the basis of actual DNA conformation, steric ion conditions, and self absorption of fluorescence by the fluorochrome. There is also the possibility of nonlinearity of the flow cytometer. However, the simplest and most probable reason for the shift resides in the nature of fluorescence itself (6).

Let us construct a simplified model of a cell (see Figure 7.5-1). The relationship between light intensity (I) and distance (x) is given by the formula

$$Ix = Io * 10^{-ecx} \qquad (7.5\text{-}1)$$

where Io is the initial intensity, e is the molar extinction coefficient, and c is the concentration of absorbing material. Transmittance (T) is defined as

$$T = \frac{I1}{Io} = 10^{-ecl} \qquad (7.5\text{-}2).$$

Absorbance (A) is related to transmittance (T) by the formula

$$A = -\log(T) = ecl \qquad (7.5\text{-}3)$$

This derivation shows that absorbance is directly related to the concentration of the absorbing material. If the absorbing material has the added virtue of being fluorescent, then what is the theoretical relationship between the fluorochrome concentration and the emitted fluorescence? Figure 7.5-2 again shows the simplified model of a cell. What we wish to do is to determine the amount of light absorbed in a narrow window through the cell, dx, calculate its resultant fluorescence, dF, and then integrate over the length of the cell to determine the total fluorescence. The light intensity in dx is given by the differential equation

$$dI = -Io*ec*10^{-ecx}\, dx \qquad (7.5\text{-}4)$$

If the quantum yield is represented by Q, then the differential fluorescence in dx is given by

$$dF = -Io*Q*ecl*10^{-ecx}\, dx \qquad (7.5\text{-}5)$$

If we now integrate over the length of the cell (l), we obtain the following formula for total fluorescence.

$$F = \frac{Io*Q}{2.3} * (1 - 10^{-ecl}) \quad (7.6\text{-}6)$$

This formula demonstrates the nonlinearity between fluorescence and stain concentration. Since the stain concentration is presumatly directly related to DNA content and fluorescent intensity is directly related to channel number, one would expect a shift to the left from ideality as stain concentration increases. For example, the observed G2-M position may be found shifted to the left of the expected 4C position. Figure 7.5-3 compares the ideal case to the actual relationship between stain content and fluorescent intensity. If ecl is much less than 1 (e.g. <.04), then the following approximation is valid

$$10^{-ecl} \approx 1 - 2.3ecl$$

This equation is the first two terms in the Taylor expansion series of an exponential. This then linearizes equation 7.5-6 to

$$F = Io*Q*ecl \quad (7.5\text{-}7).$$

There are two major points to be made about this analysis: 1) DNA histogram analysis models should not be constrained to have a 2C-4C relationship between the G0-G1 and G2-M mean positions and 2) a smaller ecl value give results approaching the ideal case.

Figure 7.5-1: Simplified model of cellular absorbance.

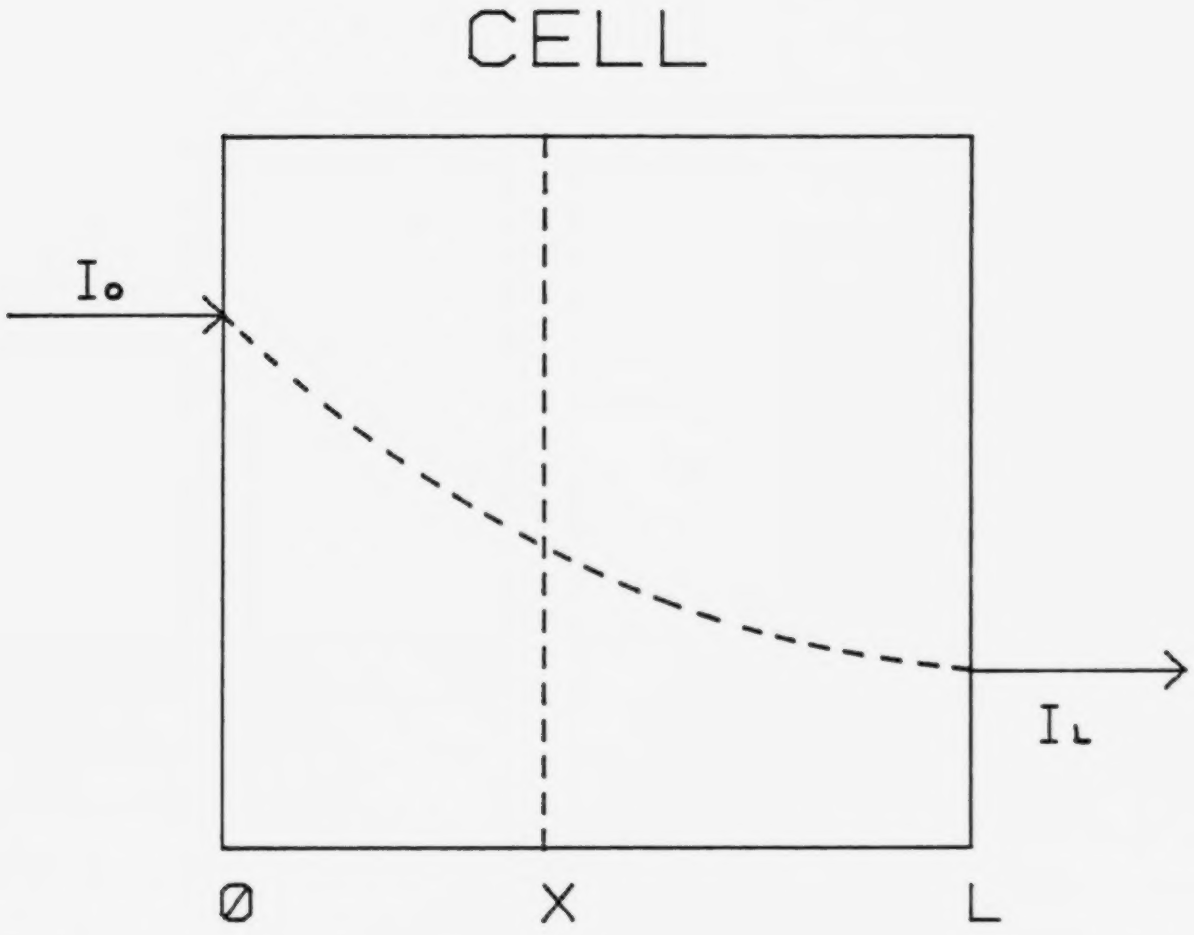

CELL
I₀
I_L
0
X
L

Figure 7.5-2: Simplified model of cellular fluorescence.

CELL

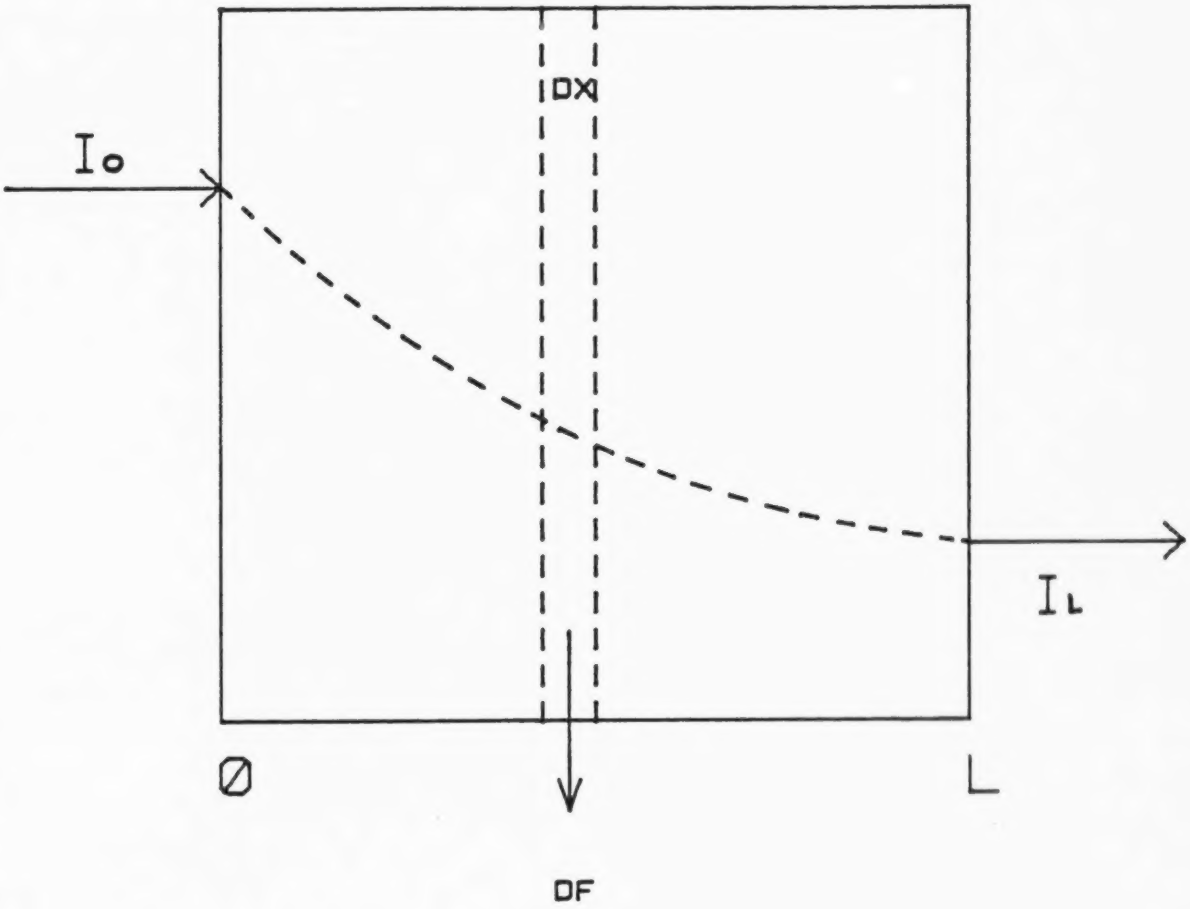
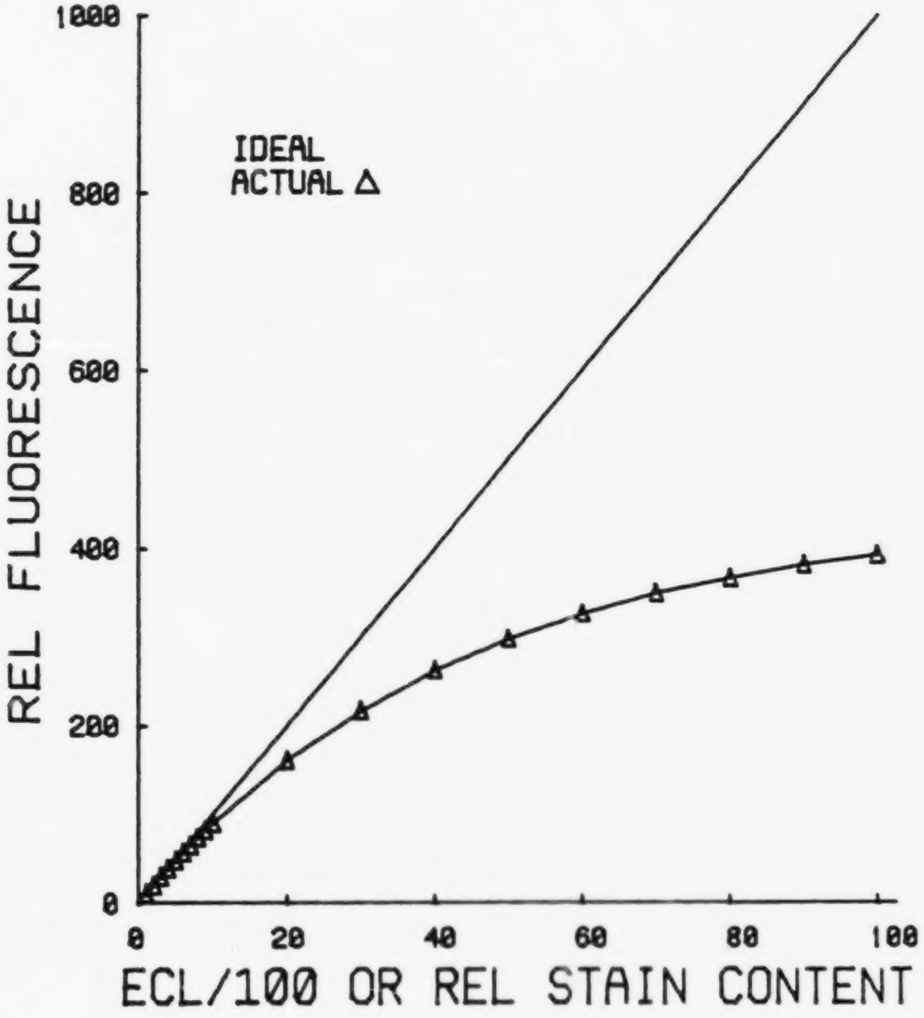

DX
I_0
I_L
0
L
DF

Figure 7.5-3: Fluorescent intensity vs stain content.


IDEAL
ACTUAL Δ
REL FLUORESCENCE
1000
800
600
400
200
0
0
20
40
60
80
100
ECL/100 OR REL STAIN CONTENT

7.6 DNA histogram kernel functions

7.61 Introduction.

There are currently two types of kernels used for DNA histogram analysis: constrained and unconstrained S phase models. The constrained method assumes S phase can be approximated by some simple function. The first and one of the most popular techniques was proposed by Dean and Jett in 1974 (28). They assumed that the kernel was two spikes (G0-G1, G2-M) flanking a second degree polynomial (S; See Figure 7.6-1). The method by which they convolved this kernel with noise was to assume that CV was constant throughout G0-G1 and S and allow the standard deviation of G2-M to be a free parameter. The total number of parameters describing the Dean and Jett model was fixed at nine (G0-G1 and G2-M peak height, standard deviation, and mean positions (6) and the three coefficients of the 2nd degree polynomial).

This method yielded excellent results with asynchronous cultures; however, when cells were perturbed, the method would often fail to give reasonable results. The problem was that in some perturbed systems, the second degree polynomial does not have suitable flexibility. One could theoretically increase the degree of the polynomial, but this causes additional problems. As the degree of a polynomial increases, the uncertainty with the fitted end points increases drastically. Unfortunately, this is the precise area where the model should be well posed; for it is in this region that G0-G1 and G2-M are being dissected away from S phase. Another problem is the determination of exactly what degree polynomial to use. The user could guess at the degree but this could add human bias to the analysis and therefore should be avoided. It is the usual practice that when faced with a perturbed population to use a kernel that has an unconstrained S phase.

In 1975 J Fried proposed a model that had an unconstrained S phase (29,30). He assumed that the kernel was just a series of spikes (See Figure 7.6-2). These spikes became Gaussian when noise was introduced into the model. This type of model was unconstrained in S since the peak heights of each of these Gaussians was a free parameter. Fried originally assumed that CV was constant and arbitrarily determined the number of S phase compartments. If he decided on too many compartments, some of the peak heights would end up being less than zero. If this occurred, he would omit that particular Gaussian and run the program again. Thus, by a trial and error type of approach, he would end up with only positive, biologically meaningful, S phase compartments. Recently Fried has made two very important improvements on this technique. He found empirically that an optimum number of compartments could be determined by the CV. He can now fix the number of compartments on the basis of an estimation of the CV. He also makes the step size between the spikes increase linearly with standard deviation (originally suggested by J Jett (31)). This increasing step size gives a much more uniform fit of the DNA histogram since the information density generally decreases at high channel numbers (Chapter 5). One of the limitations of this technique is that the ratio between the mean position and standard deviation of G0-G1 and G2-M was fixed. Originally, Fried had this ratio set at two; however, there were such a large number of cases where this did not give reasonable results when compared to autoradiography, that he made this ratio a user input. The problem with all user controlled computer routines is the possible introduction of human bias in the analysis. The major limitation of this technique was first realized by Jett. Because Fried has a discontinuous S phase, cells in very early S and late S will be picked up by the respective G0-G1 and G2-M portions of the model (See Figure 7.6-3). This causes a bias toward underestimating the S phase compartment and overestimating the G0-G1 and G2-M compartments.

In 1977 Jett published a technique (31) that attempted to extrapolate underneath the G0-G1 and G2-M compartments. The proposed kernel is shown in Figure 7.6-4. The first and last two adjacent S phase compartments are used to linearly extrapolate underneath the G0-G1 and G2-M peaks respectively. The advantages are that the S phase portion of the model is now defined at the boundaries and therefore is less apt to underestimate S phase. The problem with the technique lies with the assumption that S phase is linear for the first and last three compartments. For some cases this assumption may not be valid.

It occured to me that all these problems could be eased by simply having an unconstrained model with a continuous S phase. The simplest model that would have these attributes is the multiple rectangle kernel (See Figure 7.6-5). This model has an unconstrained S phase because the rectangles can freely move up and down to fit the data and the model has a continuous S phase between the G0-G1 and G2-M compartments. In the next chapter I will discuss this kernel and its application in more depth.

Figure 7.6-1: Dean and jett kernel function.

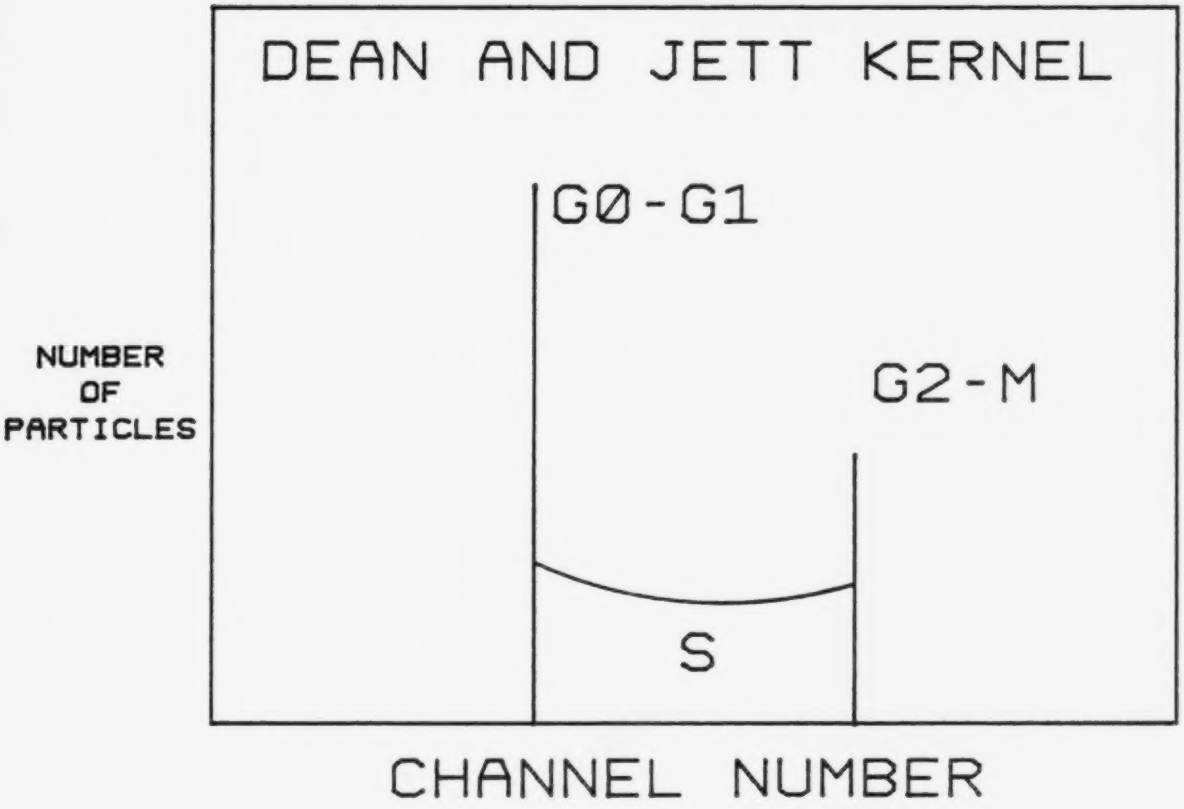

DEAN AND JETT KERNEL
G0-G1
G2-M
S
NUMBER OF PARTICLES
CHANNEL NUMBER

Figure 7.6-2: Fried kernel function.

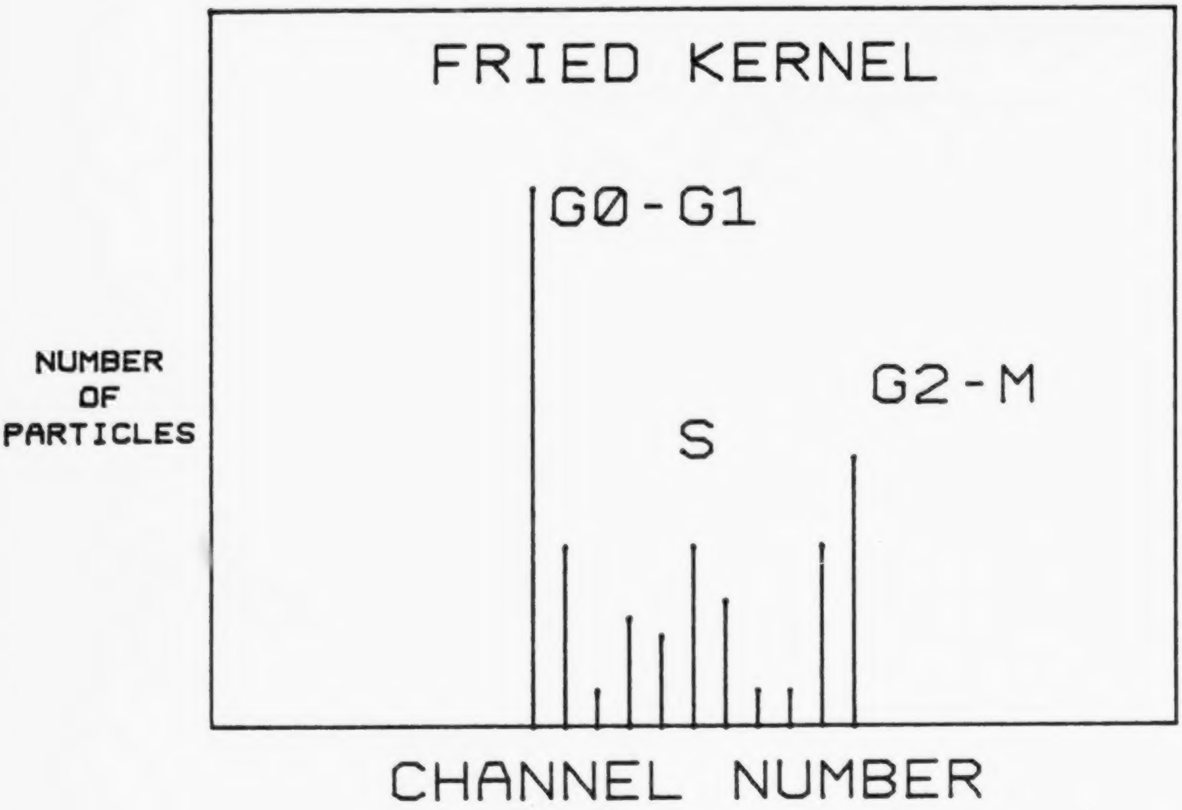

FRIED KERNEL
G0-G1
S
G2-M
NUMBER OF PARTICLES
CHANNEL NUMBER

Figure 7.6-3: Fundamental limitation of the Fried kernel function.

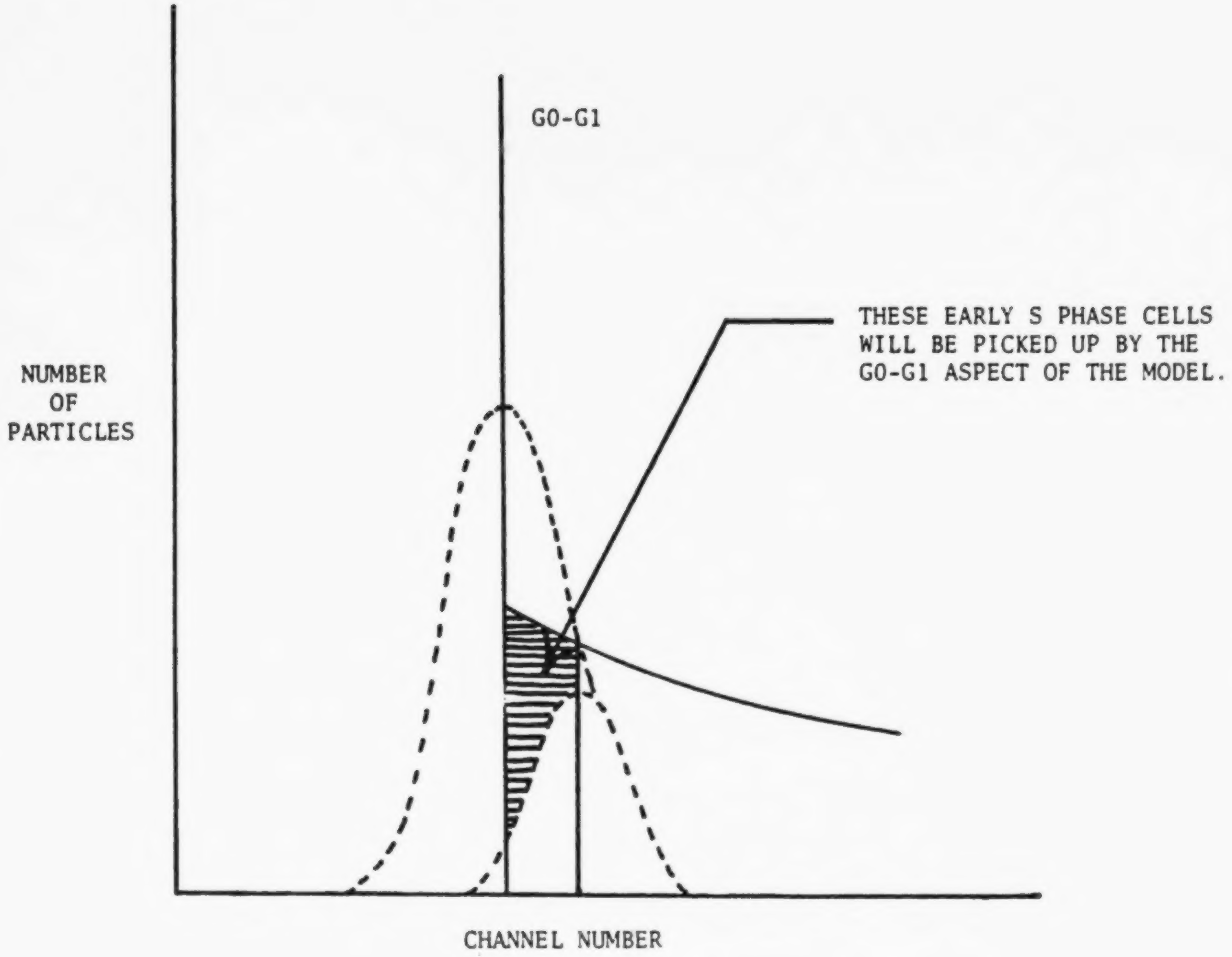

G0-G1
THESE EARLY S PHASE CELLS
WILL BE PICKED UP BY THE
G0-G1 ASPECT OF THE MODEL.
NUMBER
OF
PARTICLES
CHANNEL NUMBER

Figure 7.6-4: Jett modified Fried kernel function.

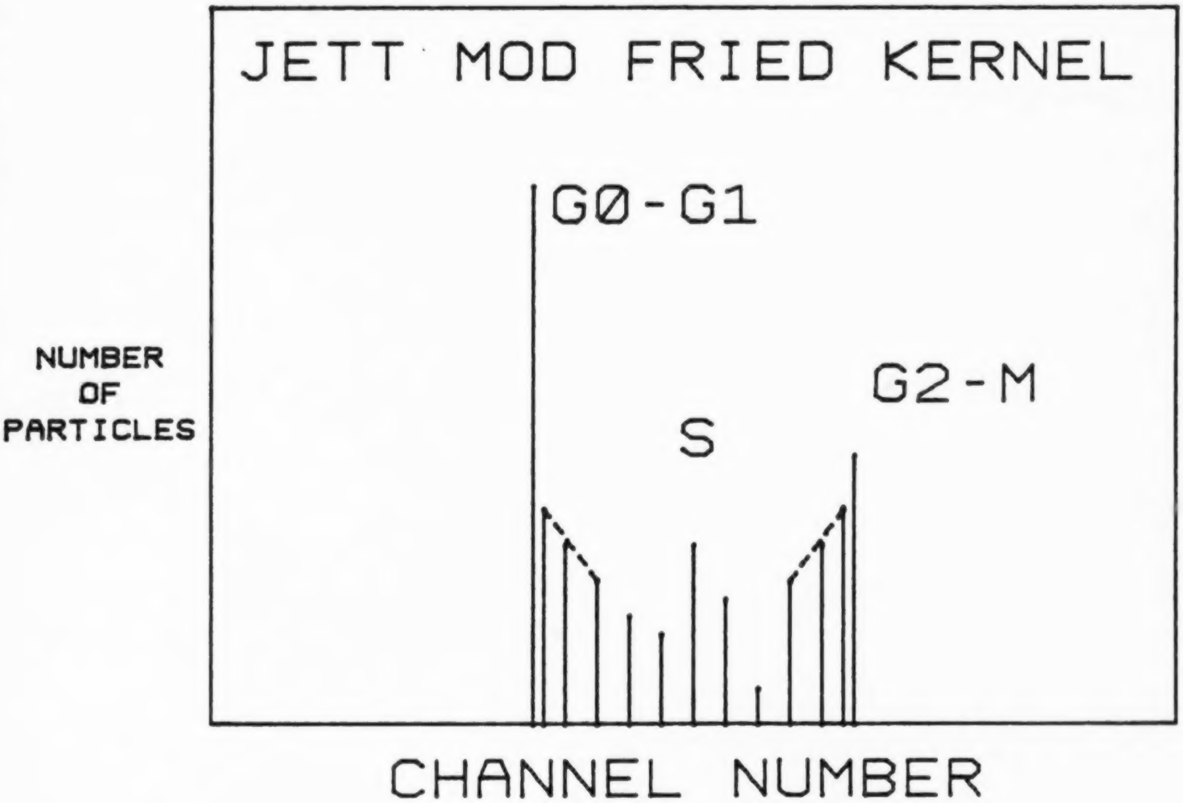
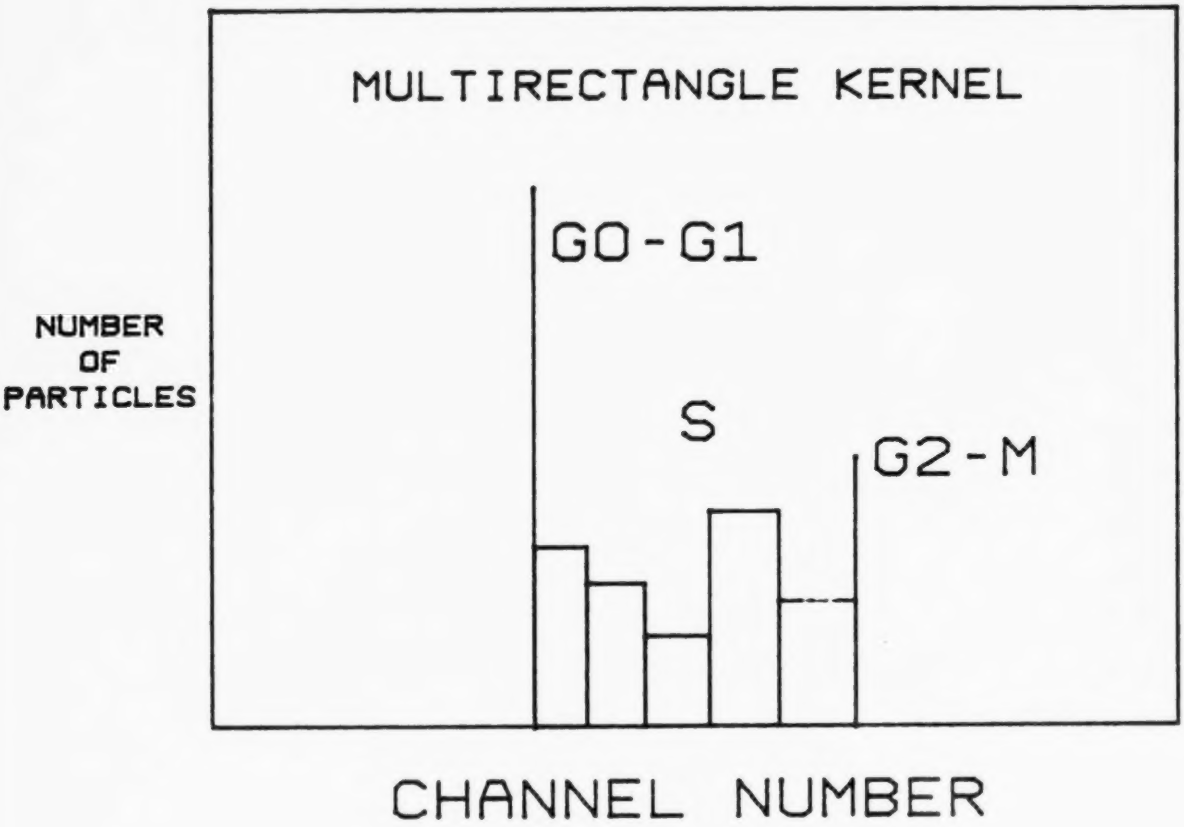

JETT MOD FRIED KERNEL
G0-G1
NUMBER OF PARTICLES
S
G2-M
CHANNEL NUMBER

Figure 7.6-5: Multiple rectangle kernel function.


MULTIRECTANGLE KERNEL
G0-G1
S
G2-M
NUMBER OF PARTICLES
CHANNEL NUMBER

7.62 Multiple rectangle technique.

There are presently two options when using this technique: standard analysis and free analysis. The standard analysis assumes that CV is constant and there is a 2C-4C relationship between G0-G1 and G2-M. This analysis, however, is usually only used for histograms with insignificant G2-M peaks. The free analysis assumes that G0-G1 and G2-M are totally independent. Both the standard deviations and peak positions are free parameters. When this free type of analysis was tried with the multigaussian approach, the program failed to converge consistantly, suggesting that it was not a well posed model with this many degrees of freedom. The only assumption that is made concerning the spread function is that standard deviation varies linearly with channel number. This relation was chosen because of the results discussed in Chapter 7.4. The fact that the mean positions of G0-G1 and G2-M are independent, allows this model to analyze histograms that do not have the ideal 2C-4C relationship (See Chapter 7.5). Finally, a Gaussian distribution was assumed for noise distribution because of the considerations given in Chapter 7.3. The only complex aspect of this model is the formulation of a rectangle convolved with a Gaussian.

Let us break the process of convolution (32) down into distinct steps in order that we can better appreciate the final formula (See Figure 7.6-6). We will start with a rectangle of height, say, Y0, and a width of A located centrally at the origin (F1(X)). The left hand side of Figure 7.6-6 shows what happens to this rectangle at each step in the convolution process. On the right side of Figure 7.6-6, we have the classical Gaussian probability density function with a standard deviation of SD (F2(X)). The first step in the convolution process is to simply substitute another variable for the independent variable X. Let us call this variable T. The next step is to flip the rectangle about the Y axis by changing F1(T) to F1(-T). This, of course, makes no difference in this case since the rectangle is symmetric about the Y axis (Figure 7.6-6 A3). If we now multiply the functions and integrate over T (Figure 7.6-6 B3), we obtain a single value that is equal to the area of F2(T) between -A/2 and +A/2 times Y0. The next step is to translocate the rectangle by X1 distance (Figure 7.6-6 B4). We now multiply this function, F1(X1-T) times F2(T) and integrate as before (Figure 7.6-6 B4). This gives us another single value. We repeat this process again and again for values X2, X3,... (Figure 7.6-6 A5,6; B5,6). If we plotted the calculated values for all real values of X, we obtain a curve such as shown in Figure 7.6-6C. The specific values of X=0, X=X1, X=X2, and X=X3 are included.

This process is called convolution and is mathematically defined as (the ranges are for this example)

$$G(X) = \int_{T=X-A/2}^{T=X+A/2} F1(X-T)*F2(T)\,dT \qquad (7.6\text{-}1).$$

If we now substitute the actual functions, we have

$$G(X) = \int_{T=X-A/2}^{T=X+A/2} \frac{YO}{SQR(2*Pi)*SD} * e^{-.5*(T/SD)}\,dT \qquad (7.6\text{-}2).$$

If we let

$$U = \frac{T}{SQR(2) * SD}$$

then

$$T = U * SD * SQR(2) \text{ and}$$

$$dT = dU * SD * SQR(2).$$

Let the limits of the integration be L1 and L2, then

$$L1 = \frac{X - A/2}{SD*SQR(2)}$$

$$L2 = \frac{X + A/2}{SD*SQR(2)}$$

Therefore, equation 7.6-2 can be rewritten as

$$G(X) = \frac{YO}{SQR(Pi)} * \int_{U=L1}^{U=L2} e^{-U^{2}} \, dU \qquad (7.6\text{-}3).$$

Since the formal definition of error function (ERF) was

$$ERF(X) = \int_{T=0}^{T=X} \frac{2}{SQR(Pi)} * e^{-T^{2}} \, dT \qquad (7.6\text{-}4),$$

the convolution formula 7.6-4 could be written simpler as

$$G(X) = \frac{YO}{2} * (\; ERF(L2) - ERF(L1) \;) \qquad (7.6\text{-}5).$$

All that remained of the problem was to find an algorithm that would calculate the error function. The following approximation to ERF(X) is used by this algorithm:

$$ERF(U) = 1 - (A1*T + A2*T^{2} + A3*T^{3} + A4*T^{4} + A5*T^{5}) * e^{-U^{2}} + e1 \qquad (7.6\text{-}6)$$

where

$T = 1/(1+P*U)$
$P = .3275911$
$A1 = .254829592$
$A2 = -.284496736$
$A3 = 1.421413741$
$A4 = -1.453152027$
$A5 = 1.061405429$
$e1 \leq .00001$

The SD that was used to convolve the rectangle was the linear interpolation between the G0-G1 standard deviation and the G2-M standard deviation at the midpoint of the rectangle. The range of the convoluted rectangle was plus and minus four standard deviations from its boundaries. The G0-G1 and G2-M Gaussians had a similar four standard deviation range. The multiple rectangle technique is now defined except for two very important features: what is the optimum number of rectangles and what are their optimal widths?

Figure 7.6-6: Steps in the convolution of a rectangle with a Gaussian.

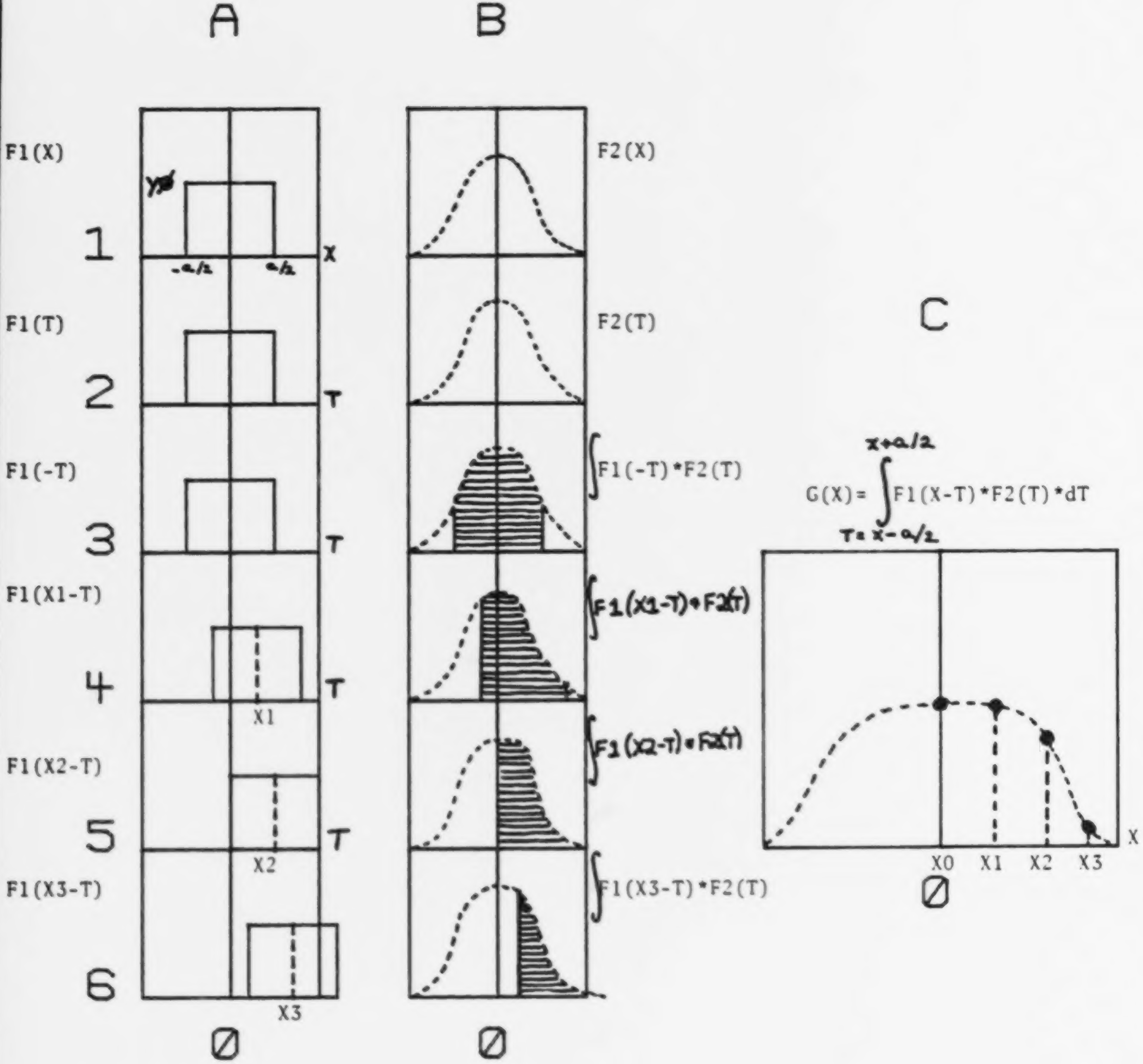
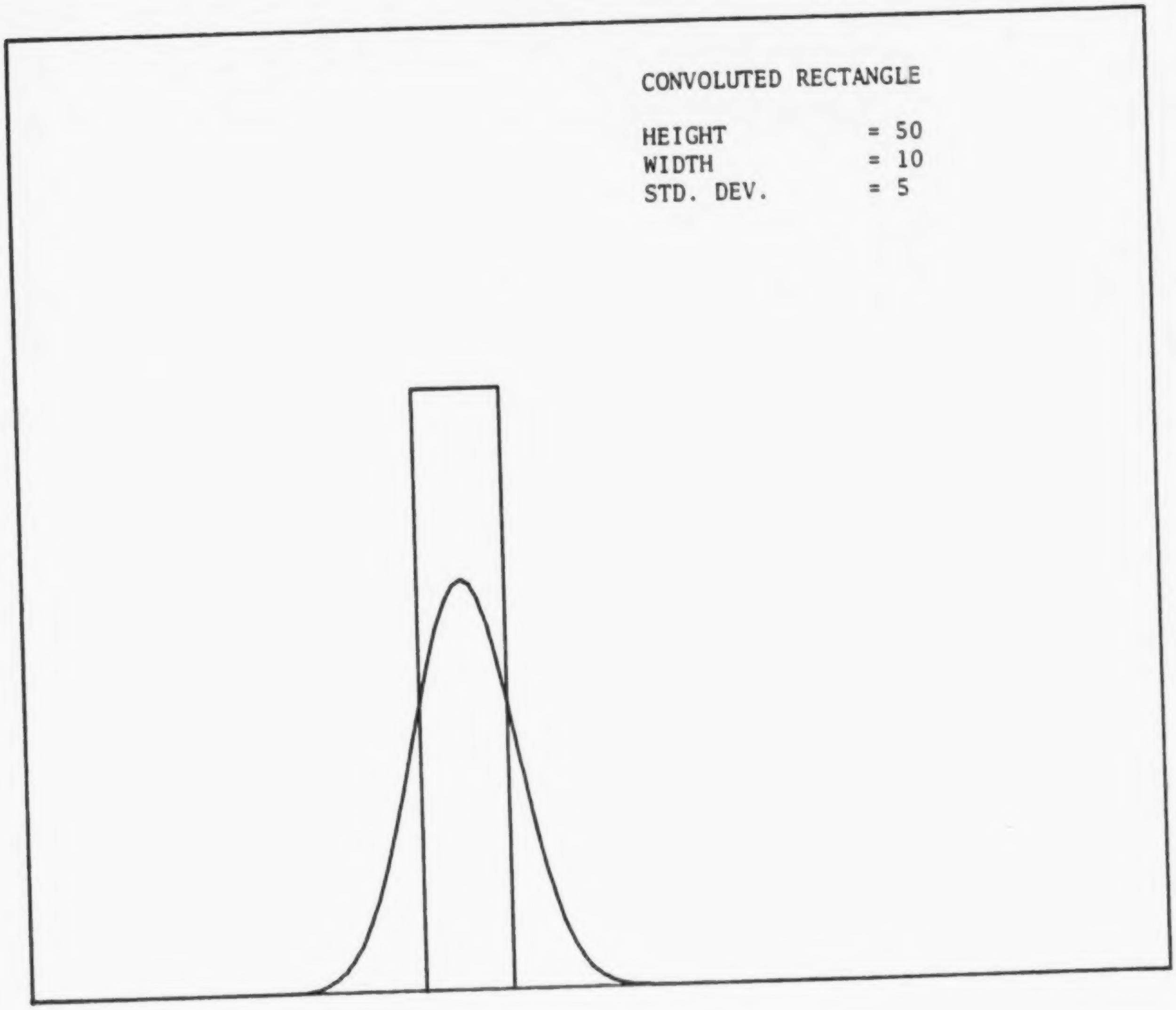

A
B
C
F1(X)
Y0
1
-a/2
a/2
X
F2(X)
F1(T)
2
T
F2(T)
F1(-T)
3
T
∫F1(-T)*F2(T)
F1(X1-T)
4
T
X1
∫F1(X1-T)*F2(T)
F1(X2-T)
5
T
X2
∫F1(X2-T)*F2(T)
F1(X3-T)
6
X3
∫F1(X3-T)*F2(T)
0
0
G(X) = ∫ from T=X-a/2 to X+a/2 F1(X-T)*F2(T)*dT
X0
X1
X2
X3
X
0

Figure 7.6-7: Convolved rectangle.


CONVOLUTED RECTANGLE
HEIGHT = 50
WIDTH = 10
STD. DEV. = 5

7.621 Optimal number of S phase compartments.

The optimal number of S phase rectangles can be computed by finding the information content of S, I1(S), using equation 5.33-9 over the estimated S phase boundaries and then determining the uncertainty associated with N rectangles with areas of S/N. The equation that determines I1(S) is

$$I1(S) = .5 * \sum_{i=G1}^{G2} \mathrm{LOG}_2 \left(\frac{Ki * S}{\mathrm{SQR}(2*PI) * CV * i} \right) \qquad (7.621\text{-}1)$$

where S is the number of S phase cells and G1 and G2 are the S phase boundaries. The uncertainty of N rectangles with a total area of S can be calculated by using an approximation to equation 5.1-7. The total number of combinations of heights of N rectangles with a total area of S is

$$C(N;S) = \frac{(S+N-1)!}{N! * (S-1)!} \qquad (7.621\text{-}2).$$

If we assume that S>>N, then equation 7.621-2 simplifies to

$$C(N;S) \approx \frac{S^{(N-1)}}{N!} \qquad (7.621\text{-}3).$$

The uncertainty associated with C(N;S) combinations is

$$H(X) = \mathrm{LOG}_2(C(N;S)) \approx (N-1)*\mathrm{LOG}_2\left(\frac{S}{N!} \right) \qquad (7.621\text{-}[illegible])$$

Since we wish to extract as much information that exists in S phase, we can calculate the optimal number of compartments by equating the information computed by equation 7.621-1 with the uncertainty calculated by equation 7.621-4.

$$(N-1) * \mathrm{LOG}_2 \left(\frac{S}{N!} \right) = I1(S) \quad (7.621\text{-}4).$$

An iterative solution to this equation is easy to implement on the computer. The value N is incremented from a starting value of 1 until the uncertainty, H(X), exceeds I1(S). The optimal number of rectangles is one less than this number. For an S phase that runs between channels 40 and 80 with 10,000 cells and a CV=5%, equation 7.621-4 computed N to be 11. In general, the computed N was slightly higher than could be obtained with the 32K memory requirements of the 4051 Tektronix. Also, the 8 bit word precision would cause undesirable effects before this value could be reached. Therefore, these calculations have limited practicality for an 8 bit microprocesor unit, but may be of interest to those with 16, 32, and 64 bit word precision.

7.622 Optimal size of S phase compartments.

The logic in determining the optimum width of the rectangles is simple if the line spread is constant. The information density under these conditions is given by equation 5.33-7 to be

$$Ici = .5*\log_2 \left(\frac{C^2 * K * Yi}{SQR(2*PI) * e * SD} \right)$$

If Yi is not rapidly changing, then Ici remains relatively constant throughout S. Therefore, the compartment sizes should also be constant. If there are to be N compartments, then the equation that determines the rectangle boundaries is given by

$$Bci = G1 + (i-1) * (G2-G1)/(N-1) \quad \text{for } i=1,N \qquad (7.622\text{-}1)$$

where G1 and G2 are the G0-G1 and G2-M positions respectively. If one has a flow cytometer with a linear line spread, a constant line spread image of the histogram can be obtained by taking the log of the intensity axis. Thus, in log space (X axis), the line spread is constant and equation 7.622-1 applies. Therefore, the mapping equation for linear line spread DNA histograms is

$$Bli = EXP(LOG(G1)+(i-1)*(LOG(G2)-LOG(G1))/(N-1)) \quad \text{for } i=1,N \qquad (7.622\text{-}2)$$

The optimal spacing for linear line spread flow cytometer has an exponentially increasing step size.

7.623 Rate and transit time calculations.

It should be intuitively obvious that in areas of high DNA synthesis rate, there will be a relative paucity of cells. This kind of phenomenon can be seen everyday during city rush hour. The chance of finding a car in a particular area of road is inversely related to its speed; thus when the velocity of traffic is allowed to increase, the average distance between the cars will also increase (assuming the road load is constant). We can take advantage of this simple principle, once we have determined the most probable kernel function from the Marquardt nonlinear least squares analysis. Dean in 1974 (13) was the first to realize that rates could be calculated from DNA histograms. His procedure, however, was designed especially for the polynomial type of DNA histogram analysis. The rate analysis that follows was designed for the multiple rectangle technique and with slight modifications can be used for the multigaussian technique.

The actual definition of DNA synthesis rate is the change in DNA content over the change in time. Since the DNA histogram does not contain any information on absolute time, the time units that must be used are relative time units. Figure 7.6-8 shows a diagram of the entire process. The X axis of the top figure is relative cell age where the cell cycle time is arbitrarily set to one. The top figure is DNA content vs relative cell age and the bottom is number of cells vs relative cell age. At the bottom of the rph is the actual calculated kernel showing the most probable noiseless compartment sizes. Our objective is to determine r1,r2,...rn in terms of known parameters. Let us start at the desired result and work backwards. The average rates for this system are

$$\begin{aligned}
r1 &= (c1-c1)/(t1-0) = 0\\
r2 &= (c2-c1)/(t2-t1)\\
r3 &= (c3-c2)/(t3-t2)\\
r4 &= (c4-c3)/(t4-t3)\\
r5 &= (c5-c4)/(t5-t4)\\
r6 &= (c5-c5)/(1-t5) = 0
\end{aligned}$$

The variables c1 through c5 are known. They are simply the boundaries of the compartments in our kernel. The only values that are not yet known are the relative transit times (t2, t3, t4, t5). The equation that describes the relationship between number of cells and relative cell age was shown in Chapter 4.2 to be given by

$$Y = 2X*\exp(-\ln2*t)$$

This equation will be more useful if the value of X is expressed in known quantities. If the total number of cells equals Z, then

$$Z-G0 = \int_{t0=0}^{t6=1} 2X * e^{-\ln 2*t}\,dt$$

$$Z-G0 = \frac{2X}{-\ln 2} * e^{-\ln 2*t}\Big|_{0}^{1}$$

After a little more algebra we have

$$Y = 2\ln 2(Z-G0)*e^{-\ln 2*t}$$

If we now integrate this equation over the G1 relative transit time interval and set it equal to the number of cells in G1, we can evaluate the first transit time, t1.

$$G1 = \int_{0}^{t1} 2\ln 2(Z-G0)e^{-\ln 2*t}\,dt$$

$$G1 = -2(Z-G0)(e^{-\ln 2*t1} - e^{-\ln 2*t0})$$

$$\text{Let}\quad K = e^{-\ln 2*t0}$$

Notice that since t0=0, K must equal one for this case. By doing a little algebraic manipulation and taking the log of both sides, we find that

$$t1 = \ln\left(K - \frac{G1}{2(Z-G0)} \right)/(-\ln 2)$$

Once we know t1, we can calculate t2 in exactly the same manner. The constant K becomes

$$K = e^{-\ln 2 * t1}$$

The relative transit time t2 is given by

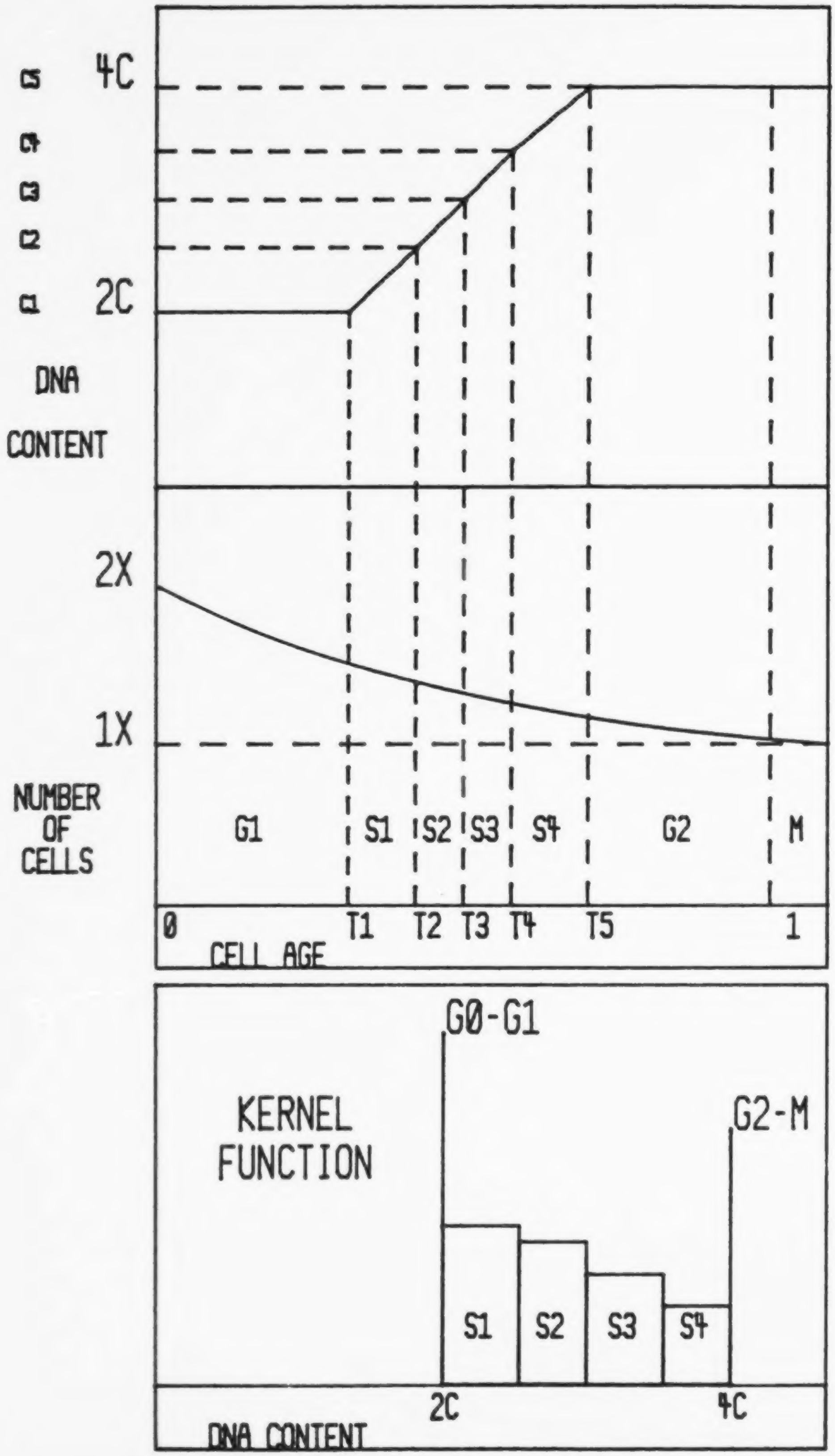
$$t2 = \ln\left(K - \frac{S1}{2(Z-G0)}\right)/(-\ln 2)$$

This process can be repeated until all the values of the relative transit times are known. These values can be directly plugged into the rate equations to give the relative rate of DNA synthesis. The positions (l1, l2, ..., l6) for each of these rates are simply

l1 = c1
l2 = c1 + (c2-c1)/2
l3 = c2 + (c3-c2)/2
l4 = c3 + (c4-c3)/2
l5 = c4 + (c5-c4)/2
l6 = c5

The generalization of this process to a variable number of compartments is obvious. A few words of caution must be said about these rate calculations. They assume that the population is asynchronous, exponentially growing, and homogeneous. Deviations from this ideal state will change the rate calculations in an unpredictable fashion.

Figure 7.6-8: Schematic of multirectangle rate calculations.


C5
C4
C3
C2
C1
4C
2C
DNA CONTENT
2X
1X
NUMBER OF CELLS
G1
S1
S2
S3
S4
G2
M
0
T1
T2
T3
T4
T5
1
CELL AGE
G0-G1
KERNEL FUNCTION
G2-M
S1
S2
S3
S4
2C
4C
DNA CONTENT

7.624 Comparison with other approaches.

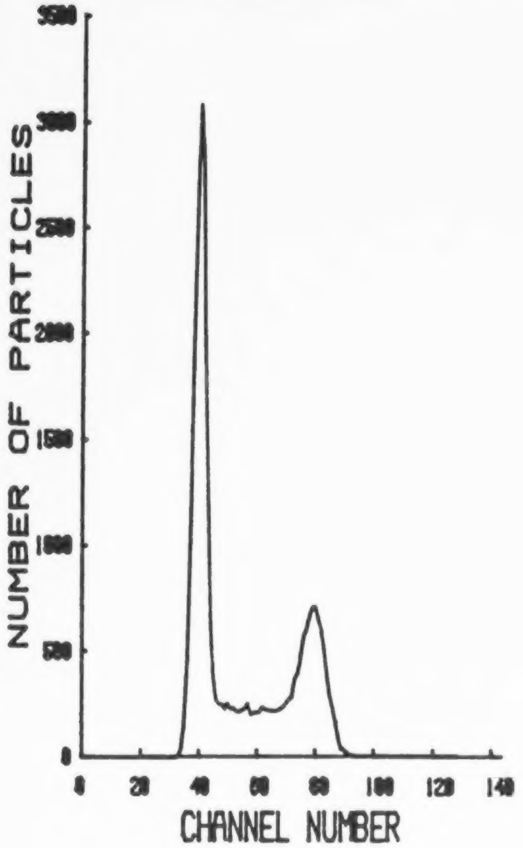
The two most popular techniques for analyzing DNA histograms are the Dean and Jett 2nd degree polynomial technique (28) and the Fried multigaussian technique (29). In order to assess whether the multirectangle technique offered any advantage over these well accepted techniques, a series of synthesized histograms were analyzed by all three methods. These histograms were designed to test various features of the analysis programs. The histograms were broken down into four catagories:

1. CV is constant, DNA synthesis rate is constant.
2. CV is constant, DNA synthesis rate is not constant.
3. CV is not constant, DNA synthesis rate is constant.
4. CV is not constant, DNA synthesis rate is not constant.

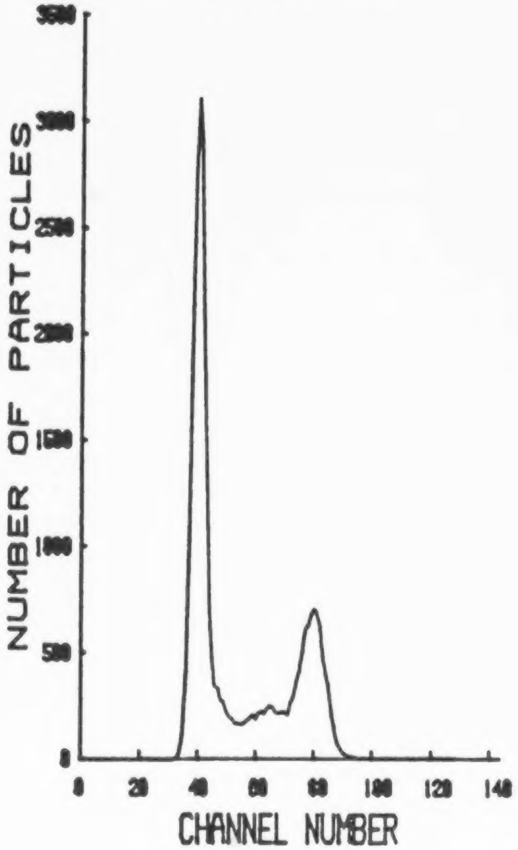
For each catagory there were two histograms: one with statistical and one without. Only the realistic histograms with statistical noise will be considered in this comparison. The histograms had the same cell cycle compartments, eg. G0-G1=50%, S=30%, and G2-M=20%, and G0-G1 CV=5. Some cell types (eg. Hela cells) have a bimodal rate of DNA synthesis (51). For the varying DNA synthesis rate groups (2 and 4), a similar bimodal distribution was chosen. The multirectangle approach was run with an 8 bit, 4051 microprocessor unit; therefore, there may be some additional uncertainty with this technique because of successive numeric truncations during the iterations. The other methods were run on a large 32 bit word system at the University of Rochester. To this author's knowledge, these programs have the same operating characteristics as those originally proposed. All three techniques allow for G2-M to float. The multirectangle and polynomial techniques have a completely unconstrained G2-M. The Fried technique assumes constant CV but allows the G2-M position to float. Both the multirectangle and multigaussian techniques were restricted to six S phase compartments. There was a minimum of user interaction with the programs. Since this comparison is based on only four histograms, a general ranking of the techniques cannot be done. However, one can appreciate general trends in the strengths and weaknesses of each technique with this data. Figure 7.624-1 shows the four histograms in the test group. The wavy appearance of the S phase for histograms 2 and 4 demonstrates the variable rate of DNA synthesis used in the generation of these histograms. Tables 7.624-1, 2, 3, and 4 show the comparisons between the three techniques. The first column is the actual values for the synthesized histogram. The 2nd, 3rd, and 4th columns are the results from the multirectangle, multigaussian, and 2nd degree polynomial techniques respectively. The first seven rows indicate the accuracy in the parameters of each model. The SSQ's gives an indication of how well each model fits the data. The lower the SSQ's value the better the fit. Finally, the number of iterations indicates the speed at which the nonlinear least squares program converged to the lowest SSQ's. The number of iterations is a function of the type of nonlinear least squares program used, the convergence criterion, and the model. Both the multirectangle and polynomial techniques use the Marquardt (CURFIT/GAUSHAUS) compromise for their nonlinear algorithms. There are a few general statements one can make about this data. Because of the theoretical limitation of the Fried technique, it will generally yield a low estimate of S (see Figure 7.6-3). This limitation can be seen in every case used in this comparison. The other general statement is that for varying DNA synthesis rates, the Dean and Jett model yielded relatively high SSQ's. This is a direct result of the constrained nature of the S phase portion of their model. Figures 7.624-2, 3, and 4 show the graphical analysis results of each technique for histogram type 2. In summary, the multirectangle technique seems to be a reasonable compromise between the Fried unconstained S phase and the Dean and Jett continuous S phase.

Figure 7.624-1: Synthesized histograms for the comparison of the multirectangle, multigaussian, and polynomial techniques. The compartment sizes are given in the test and following tables.

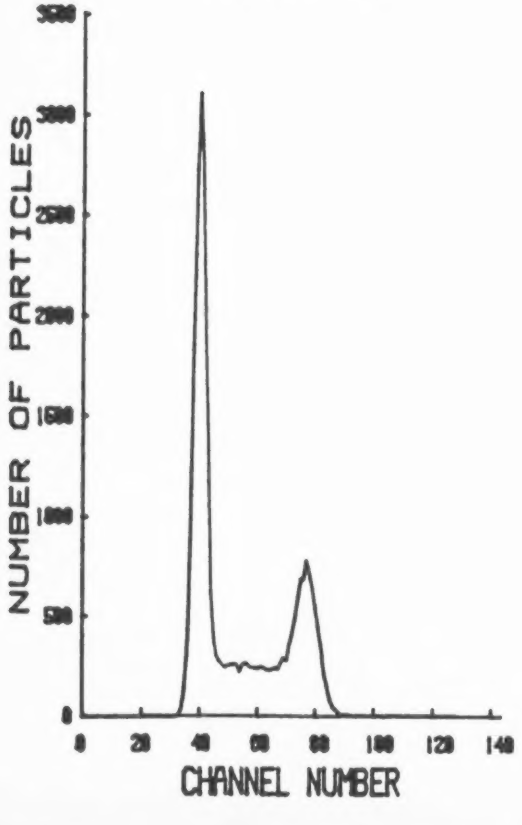
HIST 1



HIST 2



HIST 3



HIST 4

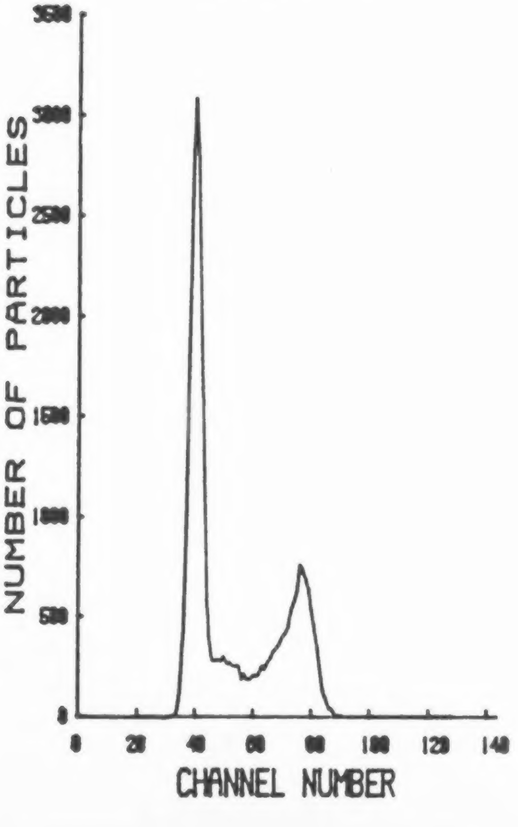


Table 7.624-1: DNA histogram comparison for test histogram 1.

HISTOGRAM 1: CV IS CONSTANT, DNA SYNTHESIS IS CONSTANT, PLUS NOISE.

	Actual value	Multirectangle technique	Multigaussian technique	Polynomial technique
%G0-G1	50.00	50.63	52.29	49.12
%S	30.00	30.01	21.32	31.79
%G2-M	20.00	19.35	26.39	19.09
G0-G1 mean	40.00	40.00	40.02	39.95
G2-M mean	80.00	80.08	80.05	80.04
SD (G0-G1)	2.00	2.02	2.02	2.00
SD (G2-M)	4.00	4.03	3.93	3.97
SSQ's	-	11,947	13,781	13,997
Number of iterations	-	4	13	5

Table 7.624-2: DNA histogram comparison for test histogram 2.

HISTOGRAM 2: CV IS CONSTANT, DNA SYNTHESIS IS NOT CONSTANT, PLUS NOISE.

	Actual value	Multirectangle technique	Multigaussian technique	Polynomial technique
%G0-G1	50.00	49.44	53.53	50.22
%S	30.00	27.64	24.04	33.83
%G2-M	20.00	22.92	22.43	15.95
G0-G1 mean	40.00	39.97	40.07	40.03
G2-M mean	80.00	79.80	79.94	80.63
SD (G0-G1)	2.00	1.99	2.01	2.01
SD (G2-M)	4.00	4.11	3.96	3.66
SSC's	-	18,474	25,323	45,589
Number of iterations	-	3	13	5

Table 7.624-3: DNA histogram comparison for test histogram 3.

HISTOGRAM 3: CV IS NOT CONSTANT, DNA SYNTHESIS IS CONSTANT, PLUS NOISE.

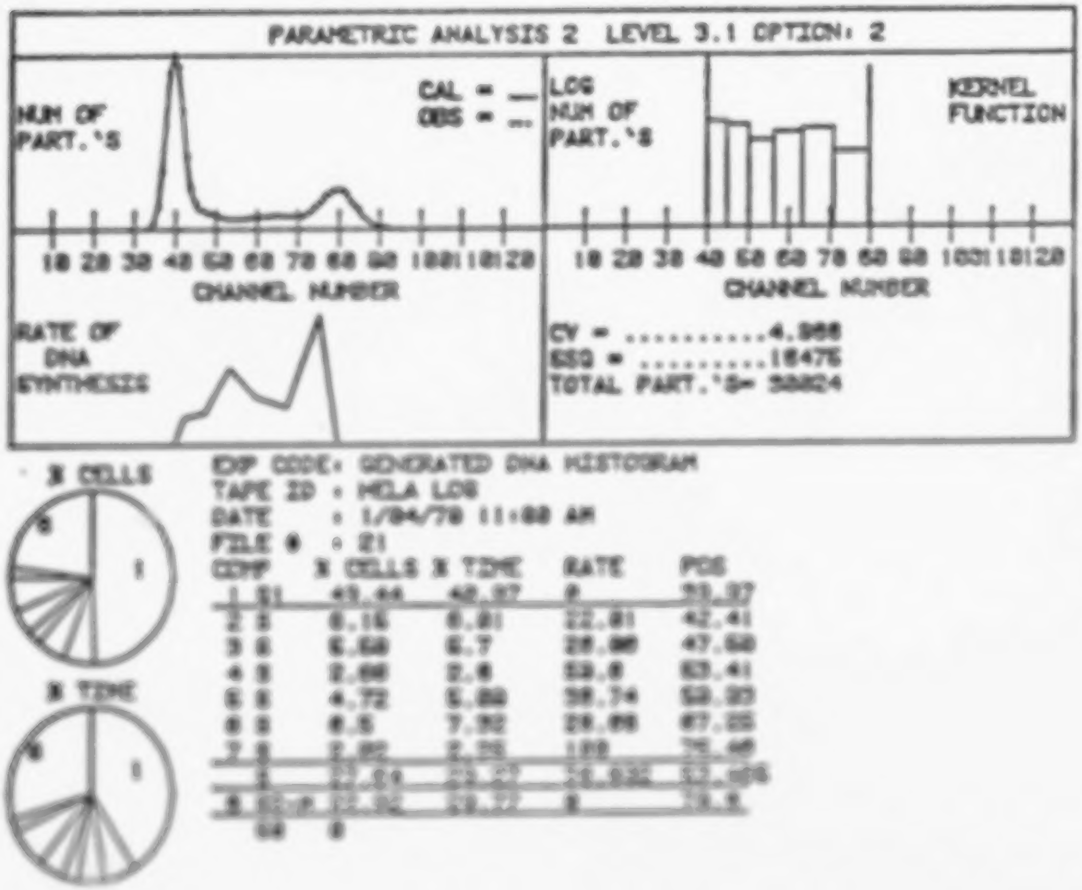
	Actual value	Multirectangle technique	Multigaussian technique	Polynomial technique
%G0-G1	50.00	49.43	52.18	49.84
%S	30.00	30.73	23.80	30.53
%G2-M	20.00	19.84	24.01	19.63
G0-G1 mean	40.00	39.98	40.05	40.02
G2-M mean	77.00	77.42	76.50	76.98
SD (G0-G1)	2.00	1.98	2.00	2.00
SD (G2-M)	3.70	3.73	4.11	3.96
SSQ's	-	14,041	26,792	31,972
Number of iterations	-	4	14	16

Table 7.624-4: DNA histogram comparison for test histogram 4.

HISTOGRAM 4: CV IS NOT CONSTANT, DNA SYNTHESIS IS NOT CONSTANT, PLUS NOISE.

	Actual value	Multirectangle technique	Multigaussian technique	Polynomial technique
%G0-G1	50.00	50.99	51.87	47.02
%S	30.00	34.68	28.92	39.79
%G2-M	20.00	14.33	19.21	13.19
G0-G1 mean	40.00	40.04	40.05	39.92
G2-M mean	77.00	78.19	77.74	78.08
SD (G0-G1)	2.00	2.02	2.00	1.94
SD (G2-M)	3.70	3.52	3.80	3.52
SSQ's	-	19,832	29,958	45,640
Number of iterations	-	5	25	9

Figure 7.624-2: Multiple rectangle graphical analysis results of test histogram 2.



```
PARAMETRIC ANALYSIS 2  LEVEL 3.1 OPTION: 2

CV = ..........4.988
SSQ = .........16475
TOTAL PART.'S= 50824

EXP CODE: GENERATED DNA HISTOGRAM
TAPE ID : HELA LOG
DATE    : 1/04/79 11:00 AM
FILE #  : 21
COMP     % CELLS % TIME   RATE     POS
 1 G1     43.44   42.37   0        33.37
 2 S      8.15    8.01    22.01    42.41
 3 S      6.68    6.7     28.09    47.68
 4 S      2.68    2.6     53.8     53.41
 5 S      4.72    5.08    38.74    58.33
 6 S      8.5     7.92    29.09    67.35
 7 S      2.85    2.55    100      75.40
   S      27.84   23.27   34.932   52.406
 8 G2+M   22.82   33.77   0        79.8
   G0     0
```

Figure 7.624-3: Multigaussian graphical analysis results of test histogram 2.

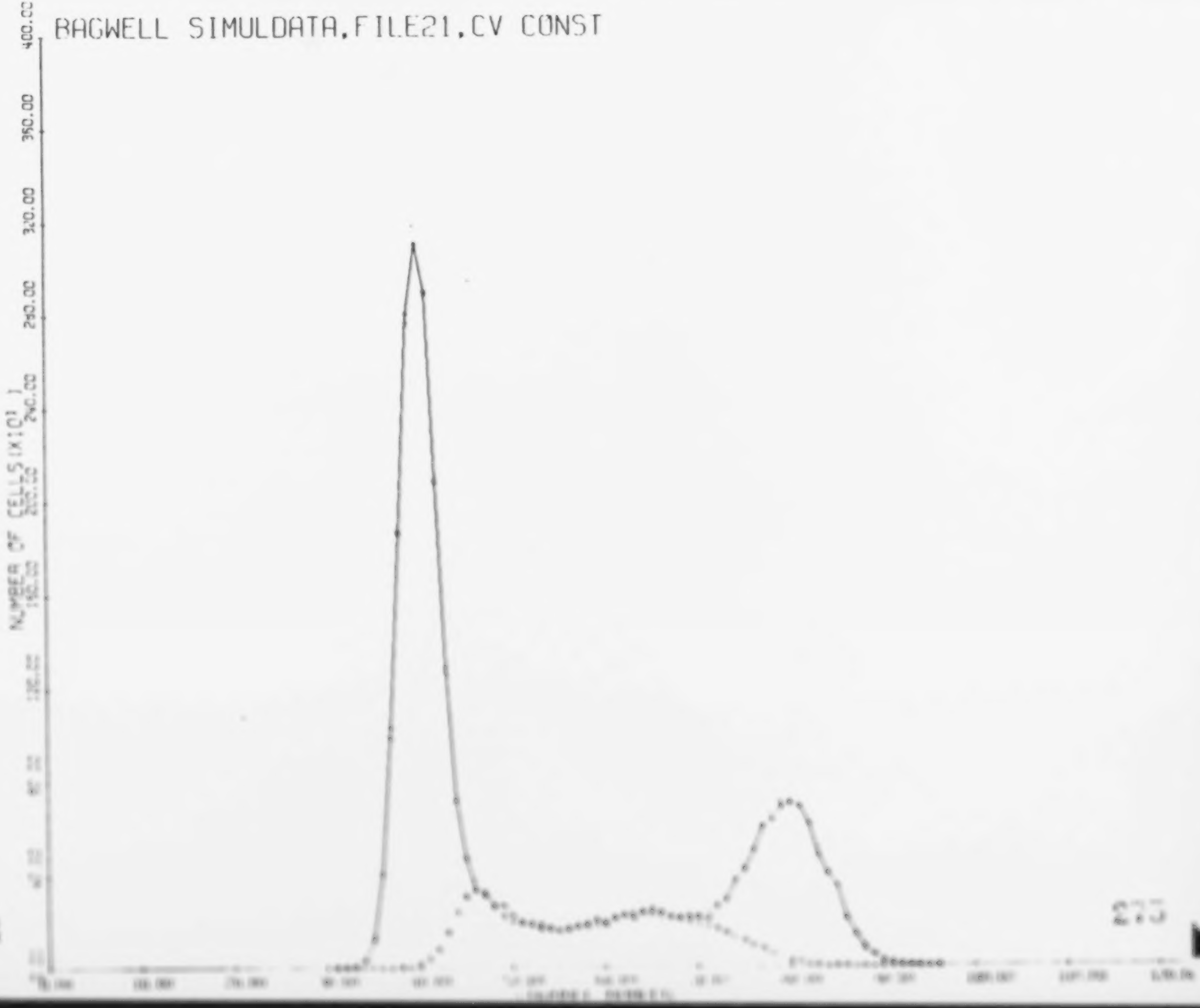
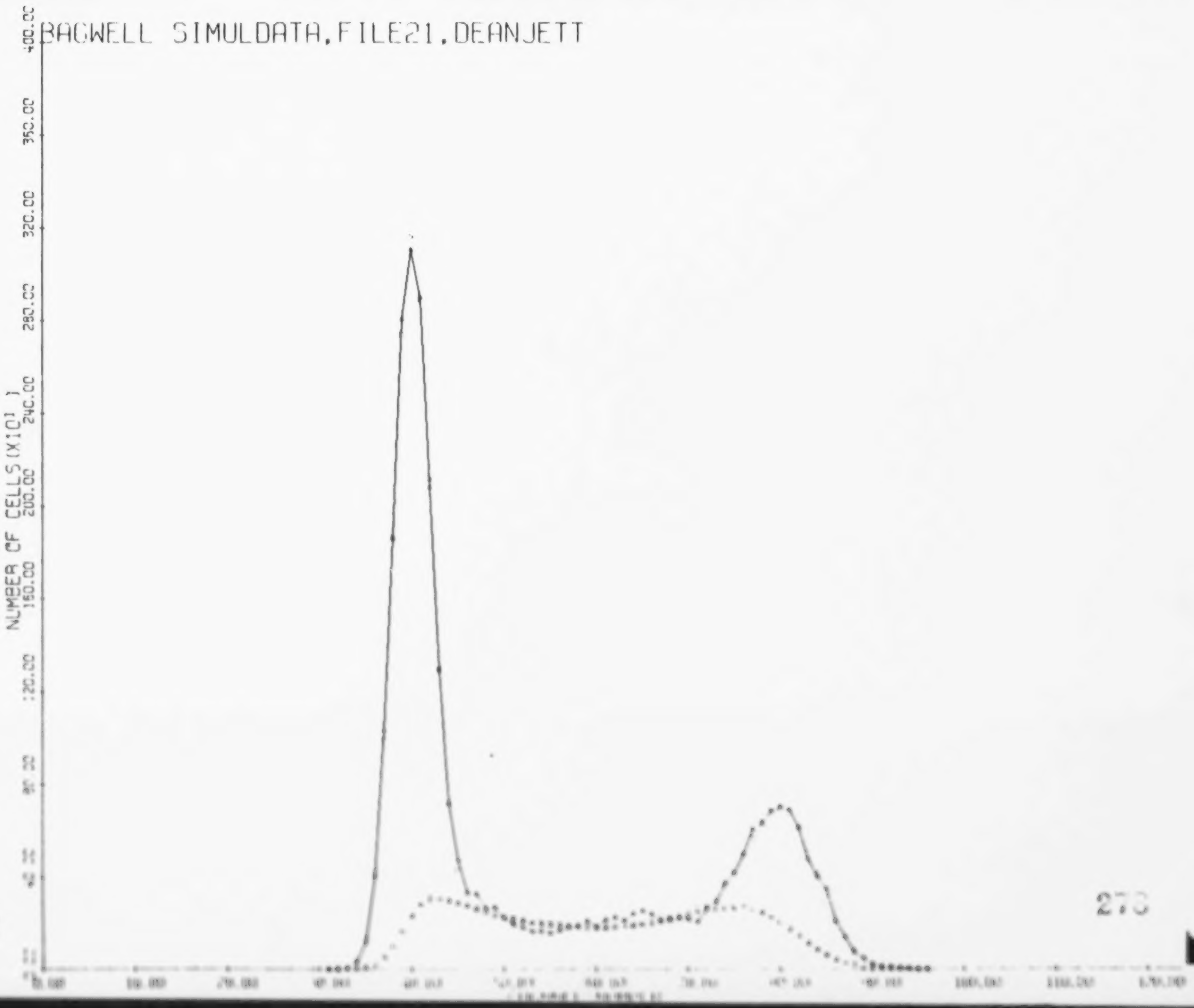

BAGWELL SIMULDATA,FILE21,CV CONST
NUMBER OF CELLS(X10^1)
400.00
360.00
320.00
280.00
240.00
200.00
160.00
120.00
80.00
40.00
0.00
10.00
20.00
100.00
110.00
120.00
CHANNEL NUMBER

Figure 7.624-b: Second degree polynomial graphical analysis results of test histogram 2.


BAGWELL SIMULDATA.FILE21.DEANJETT
NUMBER OF CELLS (X10¹)
400.00
360.00
320.00
280.00
240.00
200.00
160.00
120.00
80.00
40.00
0.00

8. Nonparametric analysis (34).

8.1 Introduction.

Flow cytometry is becoming an increasingly popular technique for measuring parameters such as cellular DNA content, RNA content, protein content, immunofluorescense, cell and nuclear size, light scatter, and fluorescence anisotropy. These parameters are usually displayed as either one or two dimensional histogram with the measured parameter(s) being on the independent axis and the number of events on the dependent axis.

One frequent problem associated with flow cytometry analysis is describing differences between one group of histograms and another group. The generally accepted method of comparison has been one of arbitrarily selecting a "typical" histogram from each set and qualitatively describing differences between the histograms. This method suffers from such serious drawbacks as selection bias, lack of information about channel to channel variability, qualitative rather than a quantitative description of differences, and lack of predictive capability.

In 1977 the Kolmogorov-Smirnov (KS) test was applied to the comparison of two different histograms (38). The test was advantageous because it made no assumptions concerning the distribution of variables or distribution of sampling error and thus was described as a nonparametric statistical analysis.

The test described in this paper is a natural extension of the 2 sample KS test. The KS test as described is only suitable for the comparison of two histograms. The method described here extends this analysis to two groups of histograms. It also is capable of finding small statistically significant areas in histogram comparisons that may have special biological meaning.

The general procedure is outlined as follows:

(1) Smooth and translocate all histograms to the same first mode channel (optional).

(2) Normalize the relevant area in each histogram to 1.

(3) Calculate the mean and standard deviation for the number of events in each channel.

(4) Perform a channel by channel statistical test such as the Student's (Gossett) t test.

(5) Select desired zone (descriptor) where there is a statistical difference between the two histograms. Use Bayes' Theorem to give the probability that the unknown histogram belongs to one of the two groups.

This procedure is nonparametric in the sense that (like the KS test) it does not attempt to reduce the data to a few parameters (eg. the number of cells in G0-G1, S, and G2-M); therefore, it can be applied to any two groups of histograms. The type of statistical test that is employed in step 4 is optional and if there are serious reservations in assuming the sampling error is normally distributed, one could use a test such as the KS test instead of the Student's t test.

8.2 Smoothing and translocation.

Smoothing the data is optional and is rarely necessary. In fact, one must be careful not to use a smoothing routine that will result in significant loss of structural information. A good review of current smoothing techniques has recently been published (18). The smoothing routine that we normally use is a variable window parabolic smoothing routine. It was specifically designed to smooth histograms less in regions of high information density (low channel numbers) than in regions of low information density (high channel numbers). The result of a typical smoothing operation on actual data is shown in Figures 8-1 and 8-2. We often use the smoothing routine because it makes our automatic peak searching algorithms much more reliable.

Also, after smoothing one can easily interpolate between points, which is necessary for histogram translocation.

In the case of DNA histograms, with a distinct G0-G1 peak we normally translocate the histograms to a first modal channel of 40. We have found that there is a small degree of shifting that occurs from one run to another. By translocating all the DNA histograms to a single position, we minimize this shifting variability. If this shifting is not compensated, there will be a large reduction in the sensitivity of this analysis.

Another technique for minimizing histogram shifting is the use of internal standards such as microspheres, chicken erythrocytes, or sorted octoploid cells. One can then use a translocation routine to move all the standard peaks to the same channel number. It is best to use a standard that has approximately the same staining properties as the sample. As seen in Figure 8-3, the original histogram (Figure 8-1) has been translocated to channel 40.

8.3 Normalize to constant area.

Once all the histograms in each series have been smoothed and translocated, they are converted to probability density functions so that sets of histograms can be averaged. In order to exclude unwanted debris or aggregation, one selects the area of interest in the histograms such as from channels 25 to 100 in our example. The normalizing algorithm determines the area from the lower boundary at channel 25 to the upper boundary at 100 and then divides every channel accumulation by the selected area. The result is a normalization of the histogram such that the area from the lower boundary to the upper boundary is equal to 1. Every histogram in the two sets is treated in the same fashion (see Figure 8-4).

8.4 Mean and standard deviation.

After all the histograms have been normalized to a constant area in the range specified, an arithmetic mean and standard deviation is found for each channel number. The locus of means for a particular data set is referred to as the mean histogram. With this procedure we form two types of histograms which we arbitrarily will refer to as type A and type B (see Figure 8-5). The variability of each data point is shown as a horizontal line representing one standard deviation above and below the mean (See Figure 8-5).

8.5 Statistical test.

The two tailed Student's (Gossett) t test (36) was used because of the relatively few samples in each set (≤10). The p values are represented in histogram form. Those p values that are greater than p=.05 are not shown and those values which are less than p=.05 are plotted on a log scale as shown in Figure 8-6. The end result is that one can visually inspect the graph for areas of statistical significance. In the example shown in Figure 8-6, there are 3 major areas of significant difference: G0-G1, late S, and G2-M. It is interesting to note that this procedure can detect small differences in S which would be diluted when comparing only percent S.

8.6 Descriptor selection and use of Bayes' theorem.

As seen at the p<.05 level in Figure 8-6, there are three statistically significant zones. The area underneath the curve in any one of these zones is referred to as a descriptor. The only restriction in selecting a descriptor zone is to stay within a region where the mean histograms do not intersect. As seen in Figure 8-6, by selecting the area under the curve from channel 86 to 95, we can then determine the probability density functions (see Figure 8-7) of the descriptor for both sets of histograms.

With unknown histograms we perform the same operations of smoothing, translocation, normalization, and descriptor selection and use Bayes' Theorem to determine the likelihood that the histogram is in class A or class B. Bayes' Theorem requires knowledge of the probability of occurence of both type A and type B. If this information is not known, it is generally assumed that the probability of occurence for each type is 0.5 (35). Table 8-1 shows the result of applying Bayes' Theorem to each member of our training set and it correctly classified each histogram.

8.7 Discussion.

The general procedure described here for determining areas of significant difference between two groups, has a number of biological applications. Probably, the most important application is the quantitation of significant changes in flow cytometric histograms associated with experimental manipulation. The data used in this paper (37) is a good example of this type of analysis. The control group of DNA histograms were from PHA stimulated human peripheral lymphocytes (HPL) and the test group of histograms were from PHA stimulated HPL treated with 200 ugs of thymocyte extract. The analysis demonstrated significant differences between control and test groups in G0-G1, late

S, and G2-M suggesting that the extract was perturbing the cycling characteristics of stimulated human peripheral lymphocytes. Because of the nonparametric nature of this test, there is no quantitation of these differences.

Nonparametric flow cytometry analysis can also be used in another type of biological application. By using a zone of significant difference between two groups as a descriptor, Bayes' Theorem can be used to classify histograms. For example, if we had a group of histograms from "normal" tissue and a group of histograms from "abnormal" tissue and a significant difference was found between the two groups, Bayes' Theorem could be used to classify unknown histograms in either "normal" or "abnormal" categories. There are a number of clinical situations where this diagnostic capability would be advantageous.

This procedure can be modified in several ways depending on the investigator's needs. The program can theoretically be extended to two or more dimensions. Instead of area normalization, there would be volume normalization for two parameters. The degree of statistical significance could be indicated by density plots or color topography maps. Also Bayes' Theorem can be extended to any number of parameters. The number of features or descriptors from the two groups of histograms can theoretically be increased which may result in an enhancement of specificity for population identification.

The chief limitation of the sensitivity of this method is the variability in the coefficient of variation (CV). This variability is mostly a function of the state of the flow cytometer (eg. partial orifice block, laser tuning, dirty optics, sample flow rate, or sheath flow rate). Thus, it is always a good practice to run a standard before and during experimental runs.

Finally, the number of numeric manipulations involved in this statistical procedure is quite small for one parameter. We use a desktop microprocessor unit (Tektronix 4051) for all our work. An analysis of the size presented in this paper (4 histograms in each group) takes less than five minutes. Most of the new microprocessor units available today can effectively handle two parameter analyses. For three or more parameters, a larger scale computer is necessary.

Figure 8-1: Raw DNA histogram. DNA histogram of 90 hour culture of PHA stimulated human peripheral lymphocytes (HPL).

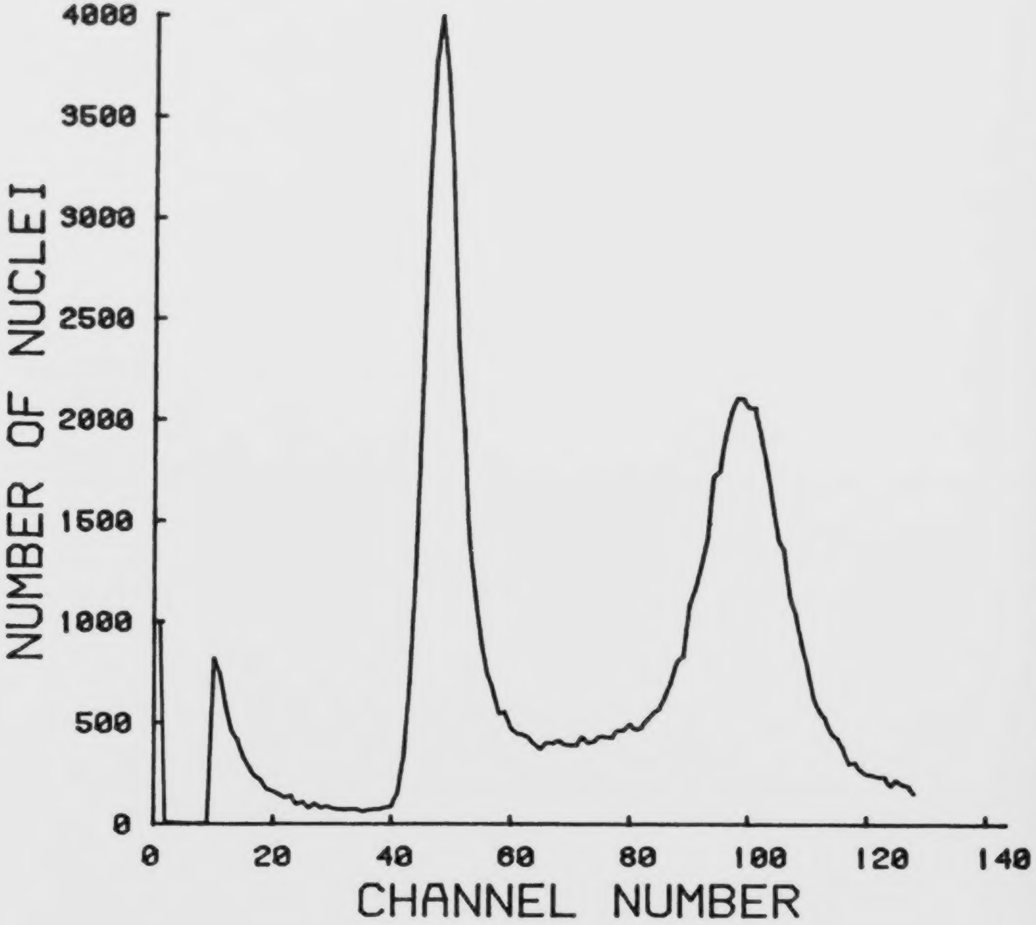

NUMBER OF NUCLEI
4000
3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100
120
140
CHANNEL NUMBER

Figure 8-2: Smoothed DNA histogram. The raw DNA histogram is smoothed with a variable window parabolic smoothing routine.

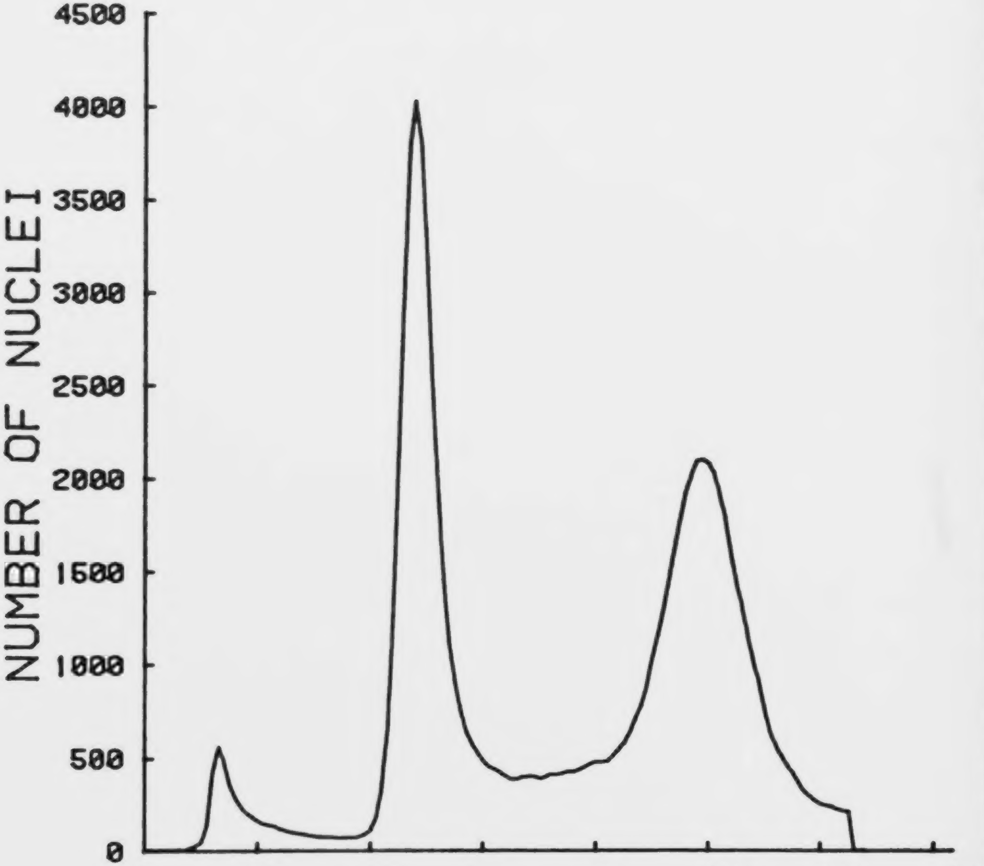

4500
4000
3500
3000
2500
2000
1500
1000
500
0
NUMBER OF NUCLEI
0
20
40
60
80
100
120
140
CHANNEL NUMBER

Figure 8-3: Translocated histogram. The smoothed histogram is mapped to a peak location of channel 40.

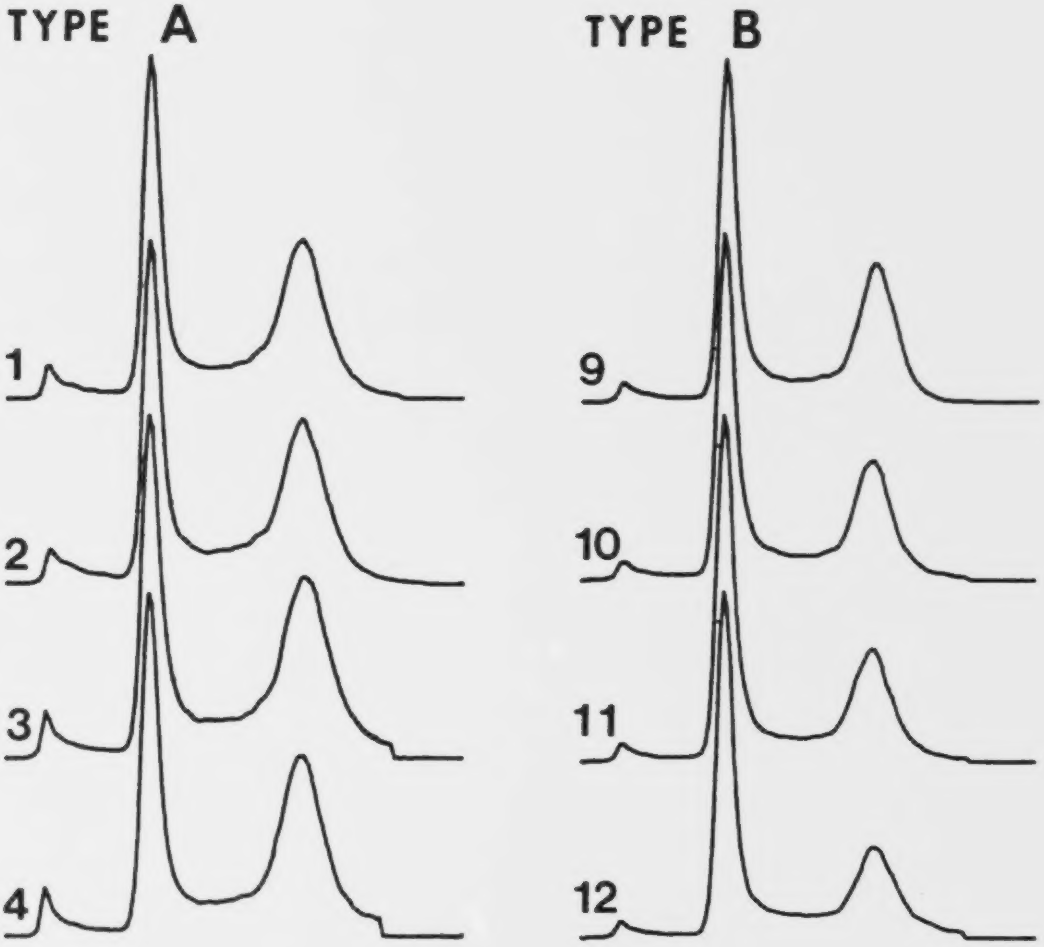

TYPE A
TYPE B
1
2
3
4
9
10
11
12

Figure 8-4: Two sets processed histograms. The first set, referred to as type A, is four DNA histograms of PHA stimulated cultures. The second set, referred to as type B, is four DNA histograms of PHA stimulated cultures with 200ugs of thymocyte extract.

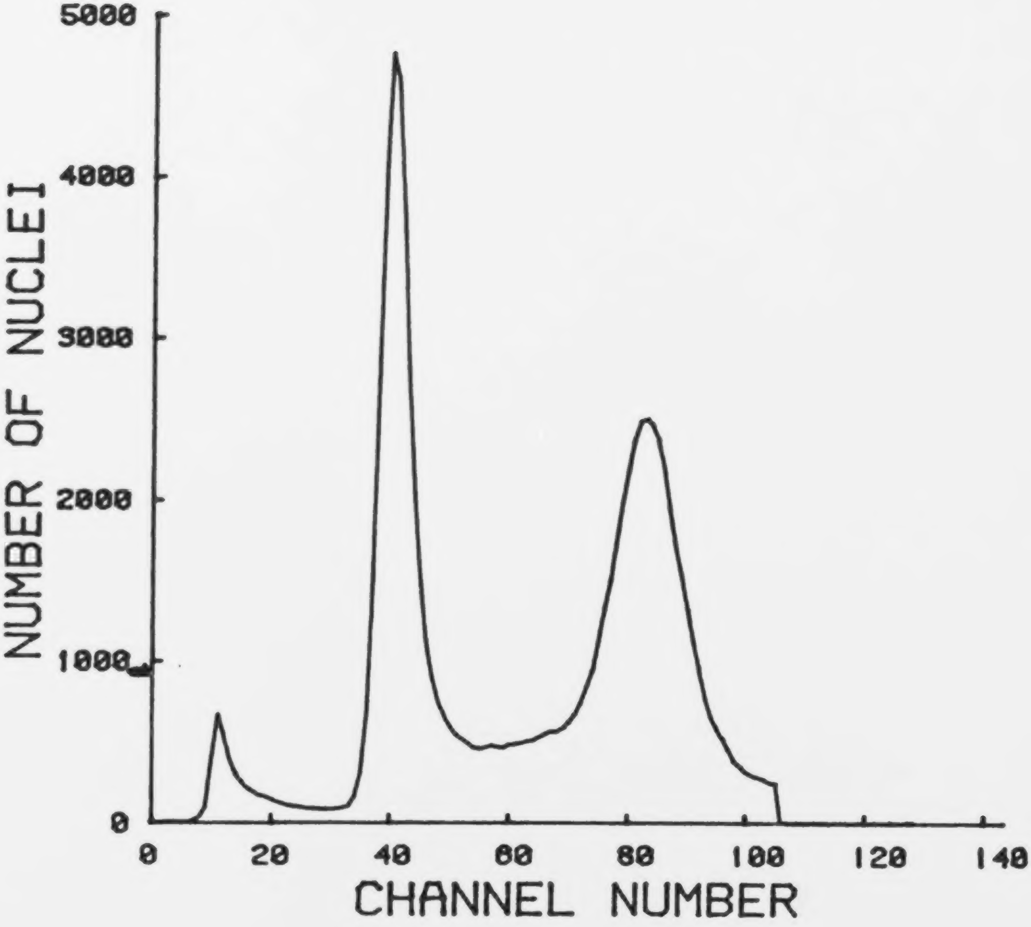

5000
4000
3000
2000
1000
0
NUMBER OF NUCLEI
0
20
40
60
80
100
120
140
CHANNEL NUMBER

Figure 8-5: Mean histograms of type A (____) and type B (----). The variability is shown as a vertical line representing one standard deviation above and below the mean. The numbers appearing below T1 and T2 refer to the computer file locations of these histograms.

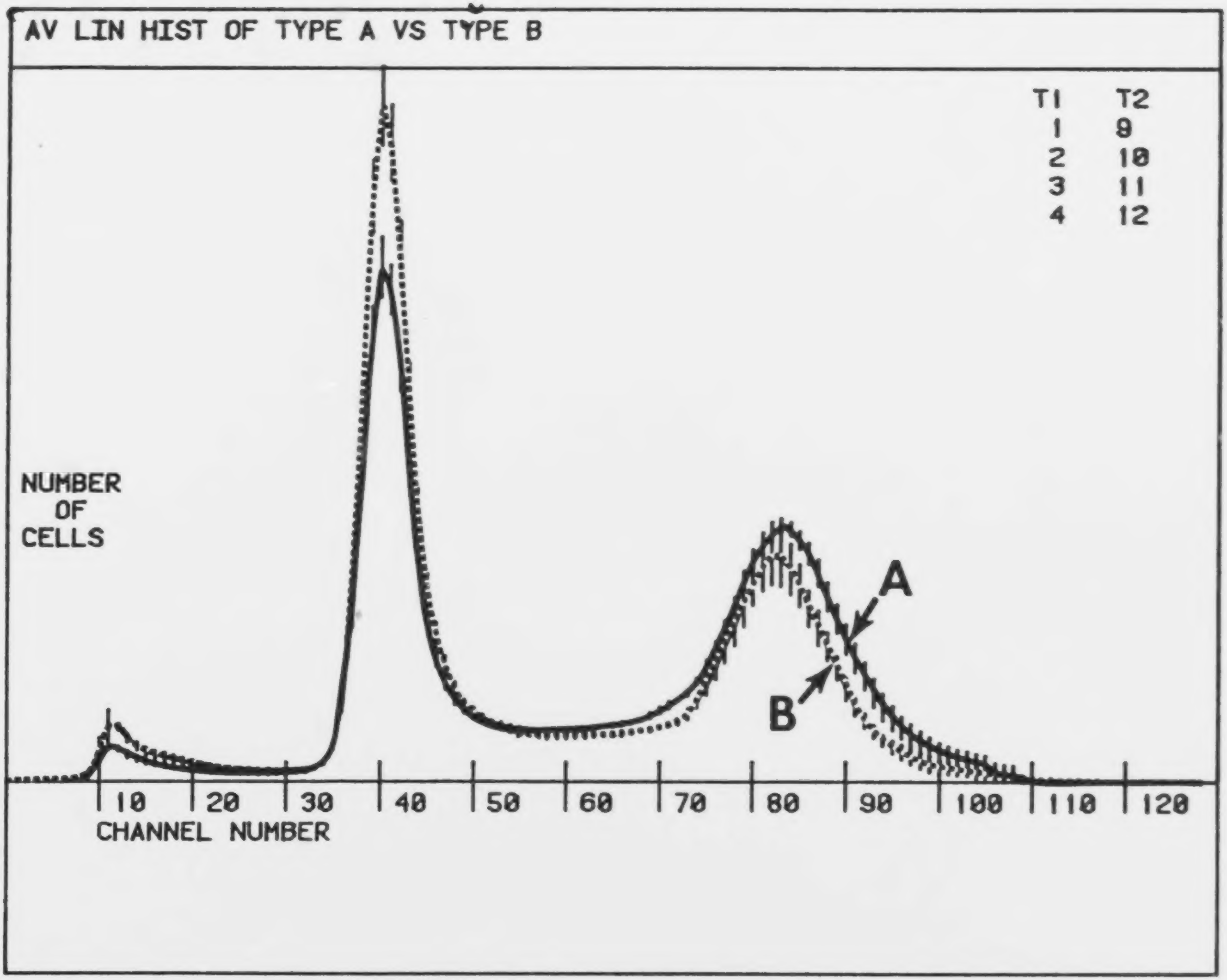

AV LIN HIST OF TYPE A VS TYPE B
T1 T2
1 9
2 10
3 11
4 12
NUMBER OF CELLS
A
B
10
20
30
40
50
60
70
80
90
100
110
120
CHANNEL NUMBER

Figure 8-6: Mean histograms of type A (____) and type B (----). The variability is shown as a vertical line representing one standard deviation above and below the mean. Channel by channel Student's (Gossett) t test is represented in histogram form at the bottom of the figure.

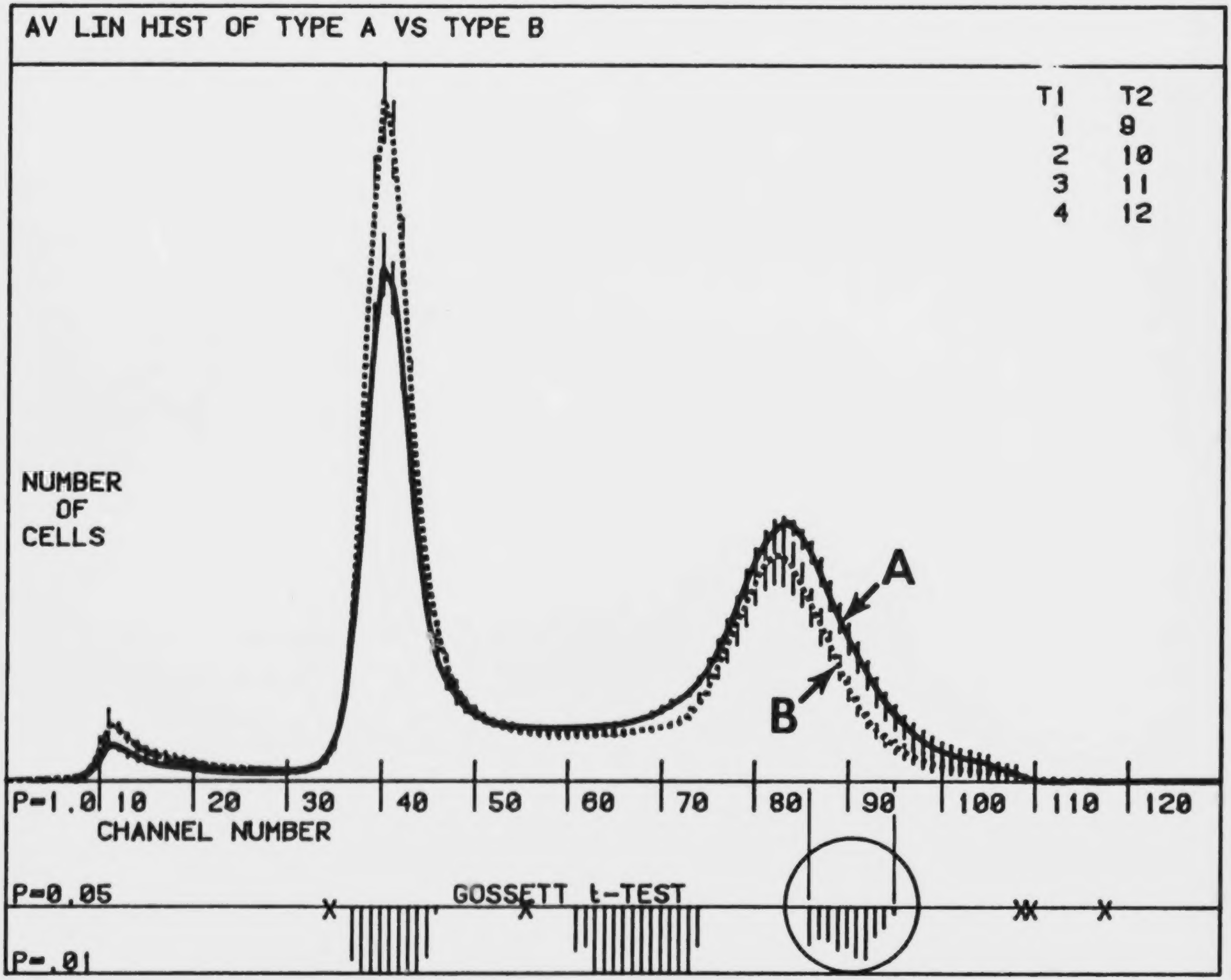

AV LIN HIST OF TYPE A VS TYPE B
T1 T2
1 9
2 10
3 11
4 12
NUMBER OF CELLS
A
B
P-1.0
10
20
30
40
50
60
70
80
90
100
110
120
CHANNEL NUMBER
P=0.05
GOSSETT t-TEST
P=.01

Figure 8-7: Probability density functions for the descriptor. Both sets of histograms were integrated from channel 86 to 95 and the mean and standard deviation of both groups is displayed as probablility density functions.

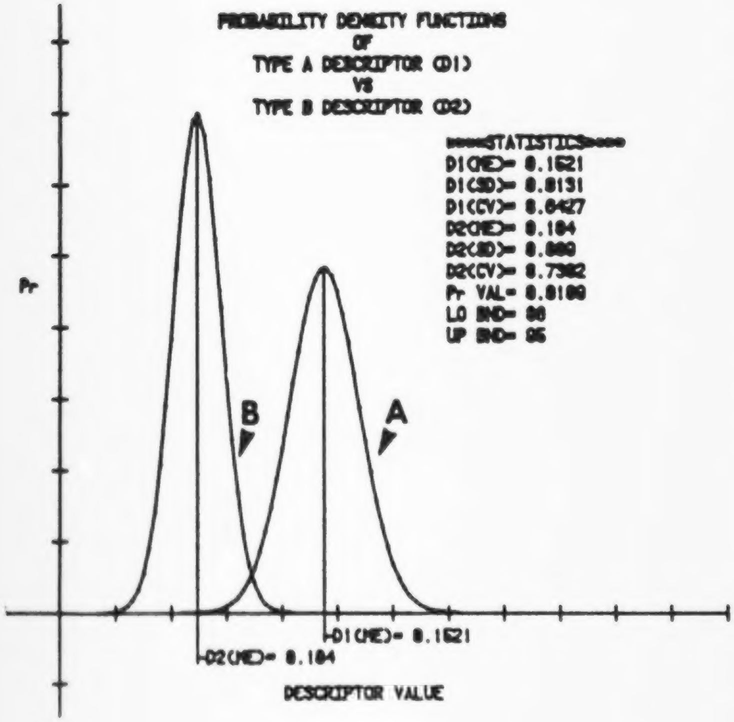

PROBABILITY DENSITY FUNCTIONS
OF
TYPE A DESCRIPTOR (D1)
VS
TYPE B DESCRIPTOR (D2)
====STATISTICS====
D1(ME)= 0.1621
D1(SD)= 0.0131
D1(CV)= 0.0427
D2(ME)= 0.104
D2(SD)= 0.089
D2(CV)= 0.7382
Pr VAL= 0.0100
LO BND= 00
UP BND= 05
Pr
B
A
D1(ME)= 0.1621
D2(ME)= 0.104
DESCRIPTOR VALUE

Table 8-1: Classification of each histogram in the training set.

```
            CHANCE OF ?? HIST. IN FILE 1 BEING CLASSIFIED AS
                 1. TYPE A = 99.99%
                 2. TYPE B = 0%

            CHANCE OF ?? HIST. IN FILE 2 BEING CLASSIFIED AS
                 1. TYPE A = 99.91%
                 2. TYPE B = 0.08%
      A
            CHANCE OF ?? HIST. IN FILE 3 BEING CLASSIFIED AS
                 1. TYPE A = 99.99%
                 2. TYPE B = 0%

            CHANCE OF ?? HIST. IN FILE 4 BEING CLASSIFIED AS
                 1. TYPE A = 99.99%
                 2. TYPE B = 0%
   ______________________________________________________________
            CHANCE OF ?? HIST. IN FILE 9 BEING CLASSIFIED AS
                 1. TYPE A = 3.05%
                 2. TYPE B = 96.94%

            CHANCE OF ?? HIST. IN FILE 10 BEING CLASSIFIED AS
                 1. TYPE A = 0.01%
                 2. TYPE B = 99.98%
      B
            CHANCE OF ?? HIST. IN FILE 11 BEING CLASSIFIED AS
                 1. TYPE A = 0.01%
                 2. TYPE B = 99.98%

            CHANCE OF ?? HIST. IN FILE 12 BEING CLASSIFIED AS
                 1. TYPE A = 0.22%
                 2. TYPE B = 99.77%
```

BEST COPY AVAILABLE

9. Parathyroids

9.1 Anatomy (39)

Parathyroids are endocrine glands that are principally involved in the homeostasis of serum calcium (40). They are yellow-brown, ovoid, encapsulated glands weighing approximately 30 mg. There are generally four parathyroids, but there is significant variation in this number (see Table 9-1).

Table 9-1: Parathyroid number distribution.

Number of parathyroids
in 527 autopsy cases

Number of parathyroid glands	Number of patients	percentage of patients with this number
6	2	0.4
5	31	5.9
4	419	79.5
3	69	13.1
2	6	1.1
Total number	527	100

The location of parathyroid glands is also highly variable This variability stems from their embryology. The upper parathyroids arise from the fourth branchial pouch and descend slighly during embryologic development along with the thyroid remaining in close association with the thyroid lobes. The lower parathyroids arise from the third branchial pouch along with the thymus and descend with the thymus. Because of the distance they travel during embryogenesis, they have a wide range of distribution. The inferior parathyroids can sometimes be found as low as the mediastinum. The inferior thyroid artery usually supplies these glands even in their abnormal positions. The surgeon will often take advantage of this anatomical fact when searching for the inferior parathyroids.

Parathyroids are principally composed of chief cells and adipose tissue. There are, however, two additional forms identified; the clear (wasserhelle) cell and the pink granular oxyphil cell.

9.2 Physiology

Hypocalcemia results in increased parathyroid hormone (PTH) secretion from parathyroid glands (PTG). PTH has essentially three target organs; kidney, intestine, and bone. The effect of PTH on the kidney is two fold; it causes decreased renal clearance of calcium and increased conversion of $1,25-(OH)_2D3$ from 25-OH D3. Recent studies suggest that 1,25-(OH)2D3 causes increased intestinal absorption of calcium. In the presence of 1,25-(OH)2D3, PTH causes increased resorption of bone. All these mechanisms tend to increase serum calcium in the face of hypocalcemia. This system will also operate in the reverse direction with hypercalemia with one major addition. There is an increase secretion of the polypeptide, calcitonin, by the C cells in human thyroid. Calcitonin directly inhibits the resorption of bone and thereby lowers serum calcium.

Thus, calcium homeostasis is a delicate feedback mechanism that precisely controls the serum calcium levels to rather small tolerances. Occasionally, a parathyroid will begin growing and producing excessive PTH. If the reason for this "stimulation" is known the patient is diagnosed as having secondary hyperparathyroidism. If the etiology is unknown, then the diagnosis of primary hyperparathyroidism or hypercalcemic hyperparathyroism is given.

9.3 Primary hyperparathyroidism

9.31 Pathology

The term parathyroid adenoma is a misnomer because it is not a always a true neoplasm arising from a single cell clone. The lethal effects from a hyperplastic or adenomatous parathyroid gland is in its over production of PTH and subsequent hypercalcemia resulting in damage to many target organs. Grossly, adenomas are usually reddish brown and smooth. They can contain cysts, hemorrhage, areas of calcification, or necrosis. This point is of particular importance in the next chapter. Adenomas are generally of the chief cell type and are usually disposed in solid sheets of cells with little fat content. Hyperplastic parathyroids are considered to have the same general histologic picture as adenomas. Most patholcgists will admit to having a great deal of difficulty in distinguishing between small adenomas and hyperplastic parathyroids. This difficulty can best be appreciated by looking at the large variation of incidences of these classifications in published studies (see Table 9-2). The pathologist's inability to make an accurate and consistant tissue diagnosis on either frozen or permanent section has lead to a bitter controversy in the surgical management of primary hyperparathyroidism.

Table 9-2: Relative frequency of entities causing hyperparathyroidism.

Author	Date	Adenomas	Hyperplasia	Carcinoma	Unknown	Total
Norris	1947	100	0	0	-	312
Cope	1958	81	11	4	4	206
Black	1961	94.5	5	.5	0	385
Cope	1966	81	15	4	0	343
Strauss	1969	53	19	4	0	58
Haff	1970	41	49	0	11	47
Goldman	1971	96	3	1	0	300
Paloyan	1973	33	65	2	0	84
Wang	1975	83	14	3	0	525

9.32 Flow cytometric detection of parathyroid hyperplasia (41).

9.321 Introduction

For the past ten years there has been considerable controversy surrounding the surgical treatment of patients with primary hypercalcemic hyperparathyroidism. Many parathyroid surgeons (42,43,44) continue to favor a conservative approach to this problem by first identifying all parathyroid glands, then excising only the single, large parathyroid adenoma along with biopsy of one or more of the normal appearing glands for histological comparison. On the other hand, there are several well known surgeons (45,46) who have favored a more aggressive approach in patients with primary hyperparathyroidism. This entails not only the excision of the largest gland, usually a parathyroid adenoma, but also excision of two and a half of the normal sized parathyroid glands, i.e., a subtotal parathyroidectomy. The primary reason for these opposing surgical approaches is the inability to determine hyperfunction or hyperplasia of the smaller and histologically nondiagnostic glands remaining in the neck after excision of an obvious parathyroid adenoma.

The purpose of this study was to measure and compare intranuclear DNA content in cells from normal, hyperplastic and adenomatous parathyroid glands by the method of flow cytometry. The differences observed between normal and adenomatous parathyroid glands were used to identify histologically undetectable hyperplasia.

9.322 Materials and methods

For the past two years multiple biopsies were taken for DNA analysis from pathologically confirmed parathyroid adenomas removed from thirty-seven patients with primary hypercalcemic hyperparathyroidism. In these patients a small biopsy was taken from each of the remaining normal-sized glands that could be identified. A portion of the biopsy was histologically examined by frozen section and usually read as normal or nondiagnostic parathyroid tissue. These glands will be referred to as "normal" indicating that they came from patients with hyperparathyroidism from whom an adenoma had been removed. The diagnosis of hyperplasia was histologically present in only eight of the fifty-six glands examined. Small biopsies of normal parathyroid glands were obtained from normocalcemic patients with no known parathyroid disease. These samples were obtained from pathology specimens submitted for examination in patients with either thyroid carcinoma and suspected lymph node metastases, or from glands inadvertently damaged and transplanted during total thyroid lobectomy for possible malignant disease. Only two patients in the normal group had received prior irradiation, and most had received exogenous thyroid preparations prior to surgery for their suspected malignancies.

Each sample was teased in cold 0.1M Tris buffer (pH 7.0) and stained with propidium iodide according to the method of Vindelov (11). The nuclei ($5 X 10^5 - 1 X 10^6$ nuclei/ml) were filtered through a 30 micron nylon mesh, examined by a fluorescent microscope for doublet contamination and analyzed by the TPS-1 flow cytometer (Coulter Electronics, Inc., Hialeah, Florida) to generate a DNA histogram. Approximately 20,000 to 60,000 nuclei were counted over a five to ten minute interval. A one to ten percent doublet contamination was encountered by fluorescent microscopy due to the adhesive nature of normal parathyroid nuclei. Adenomatous tissue, however, was easily dispersed and maintained as a single cell suspension. The doublets were mathematically removed from all samples by an interpolative process involving the examination of the number of singlets and triplets (Chap 6.7). Since propidium iodide stoichiometrically intercalates with DNA, the emitted fluorescent intensity caused by laser excitation of each nuclius is proportional to DNA content. The degree of fluorescence for each nucleus is stored as single count in a multichannel analyzer. The accumulation of all these counts can be displayed as a DNA histogram with an abscissa of DNA content and an ordinate of number of nuclei. Figure 9.3-1 shows a typical DNA histogram from normal parathyroid tissue. The first peak represents G0-G1 nuclei with a diploid amount of DNA. There are virtually no nuclei in the cycling phases of S and G2+M. In contrast to this quiescent pattern, Figure 9.3-2 shows a DNA histogram from a parathyroid adenoma. The increase in tetraploid DNA can be appreciated by the number of nuclei in the G2+M phase. An octaploid peak was observed in some adenomas. In order to determine whether the tetraploid difference observed between normal and adenomatous parathyroid glands was statistically significant, a nonparametric flow cytometric analysis was implemented (Chap 8). This test calculates the mean histogram for each group and performs a channel by channel Student (Gossett) t test to find areas of significant difference. The comparison between averaged normal and adenomatous parathyroids is shown in Figure 9.3-3. The difference between the two groups in the G2+M region (channels 50-70) was signficant at the $p<.04$ level. The percent of nuclei that fall within this tetraploid region will be referred to as the hyperplastic index.

9.323 Results

The hyperplastic index for each histologically diagnosed parathyroid specimen is shown in Figure 9.3-4. The means of the hyperplastic indices for the normal and adenomatous parathyroid groups were 0.18% and 3.3% respectively ($p<.04$). The mean hyperplastic index for the glands histologically diagnosed as hyperplastic was 2.0% ($p<.02$ when compared with the normal group). In the patients with parathyroid adenomas, "normal" glands had a hyperplastic index of 0.38%. This group had a wide variation with many samples falling in the normal and adenomatous ranges.

The variability about the means for the normal and adenomatous groups can be expressed as probability density functions as shown in Figure 9.3-5. These probability distributions can be used to determine the likelihood of any unknown parathyroid sample being adenomatous based on Bayes' theorem. These probability calculations have been applied to every parathyroid biopsy in this series and the results are shown in Figure 9.3-6. The only difference between this figure and Figure 9.3-4 is that the ordinate is presented in terms of percent chance of the sample being an adenoma. The 50% line is the decision axis where samples above this line are recognized by the computer as positive for being in the adenomatous range. Those samples that fall below this line are considered to be in the normal range and therefore negative. The entire series is summarized in Figure 9.3-7. All of the biopsied normal glands were correctly identified by the computer. In the eighty samples tested from thirty-seven adenomas, sixty-eight were correctly identified (85%) and there were twelve false negatives (15%). Two of these negative samples came from a patient whose adenoma had aneuploid DNA which was undetected by our technique. Three samples from one other adenoma were all negative; however, two of the other parathyroids found in this patient were histologically hyperplastic and flow cytometrically positive. The other seven false negatives were from adenomas in which biopsies from other areas of the same adenoma were positive. Eight grossly normal appearing glands were histologically hyperplastic. Seven of these glands (87.5%) were positive with greater than 50% chance of being an adenoma and only one was negative (12.5%). In the forty-eight "normal" glands biopsied from twenty-eight hyperparathyroid patients in which an adenoma had been removed, thirty-four were classified as negative (71%) and fourteen were positive (29%). When multiple glands were biopsied, two patients had all positive glands, eight had all negative glands and the remainder were mixed.

9.324 Discussion

In our own twelve-year experience of 163 patients operated on for the first time with the diagnosis of primary hypercalcemic hyperparathyroidism, we have practiced the conservative surgical approach, namely, excision of the single large adenoma, visualization of the other three grossly normal glands when possible, and biopsy of one or more of these parathyroids for histological confirmation. In studying failure rates in this group, it is important to separate patients with persistent hypercalcemia following surgery from those with true recurrence defined by hypercalcemia after a normocalcemic postoperative interval. Failure to separate patients in this manner may explain the differences in the reported recurrence rates (44,46). Persistent hypercalcemia is usually caused by misdiagnosis or failure of the surgeon to remove all of the abnormal parathyroid tissue. In our series, there were nine patients (6%) identified with persistent hypercalcemia following their first operation. Of the 92 patients with a good normocalcemic response who were studied at least three months after cervical exploration, 7 (8%) were found to have late recurrence with total serum calcium greater than 11 mg/dl. Recently we have examined more closely "cured" patients followed from one to eleven years after parathyroidectomy with serum levels of parathyroid hormone and ionized calcium. It is interesting to note that although these patients continue to have normal total serum calcium levels (<11mg/dl), 50 percent of the group have both parathyroid hormone levels elevated to greater than 100 ulEq/ml (n≤60 ulEq/ml) and ionized calcium levels greater than 2.40 mEq/l (n≤2.37 mEq/l). Attie and co-workers reported elevated parathormone levels in 15% of their 74 parathyroidectomized patients, only one of whom was known to have an abnormal total serum calcium (47). These findings suggest that this group of patients may still be at risk for recurrent hyperparathyoidism if their life expectancy is of long duration.

Recently, more asymptomatic patients are being presented to the surgeon earlier in the course of their hyperparathyroidism. In these cases, it is sometimes difficult to determine multiglandular involvement, especially if the adenoma is not strikingly larger than the other glands.

Our pathologists often will not commit themselves as to whether a parathyroid gland is hyperplastic or normal on frozen section. The decision to excise a parathyroid gland in these patients is based mostly on the estimated gross size of the gland and the surgeon's experience. Attempts at identifying hyperplasia by methods other than the poorly defined histological criteria presently used, seem warranted to improve surgical results.

The Wang (48) density test has been proposed as a means of identifying hyperplasia in parathyroid gland biopsies at the time of surgery. This procedure has not gained wide acceptance because it only measures the fat content of the gland in question rather than funtional or proliferative activity of the parathyroid cells. Bengtsson, et al (49) using a cytofluorometric technique with parathyroid cell imprints found a difference in nuclear DNA content between adenomas and normal parathyroid glands. In contrast to adenomatous glands, these authors described no tetraploid nuclei in normal or hyperplastic glands except in one patient with secondary hyperparathyroidism. With the newer, more rapid and accurate technique of flow cytometry for measuring DNA content, we have confirmed their findings and have extended them to demonstrate the differences in normal, "normal", hyperplastic, and adenomatous parathyroids. As expected, based on Bengtsson's results, a significant difference between normal and adenomatous parathyroids was found. However, in contrast to their study, we found a significant difference between normal and hyperplastic parathyroid glands. One explanation for this difference is the increased sensitivity of flow cytometry over manual microscopic methods.

A new and interesting aspect of this study is the variation in DNA content present in "normal" glands from hyperparathyroid patients. Although the majority were found to have little tetraploid DNA, some samples had a hyperplastic index similar to hyperplastic and adenomatous glands. In the presence of a large adenoma, there would be suppression of the "normal" glands if the etiology of this disease involved only a single, autonomous parathyroid neoplasm. Since twenty-nine percent of adenomatous patients had one or more positive "normal" glands, an idiopathic stimulus causing parathyroid hyperplasia is implicated as one of the major etiologies of this disease. This concept is supported by Fialkow et al (50) who demonstrated a multicellular origin in parathyroid adenomas by utilizing the heterozygous X-linked isoenzyme marker, glucose-6-phosphate dehydrogenase. It is our hypothesis that these computer-identified positive glands may contribute to the occassional late recurrence seen in this disease. If this hypothesis is proven with clinical correlation, it will be possible to selectively identify and excise high risk parathyroids based on the hyperplastic index determined by biopsy. It is technically feasible to completely process a biopsy for flow cytometric analysis in 20-30 minutes; therefore, this technique can be performed during surgery along with a frozen section. Since seventy-one percent of the remaining glands were found to have a low hyperplastic index, the unselected excision of glands as practiced in prophylactic subtotal parathyroidectomy is unwarranted.

Figure 9.3-1: Typical DNA histogram from a normal parathyroid. The abscissa is fluorescent intensity which is proportional to DNA content. The ordinate is number of nuclei counted. The first peak represents diploid nuclei in the resting cell cycle phases of G0 and G1. The abscence of nuclei in the phases of S and G2+M should be noted.

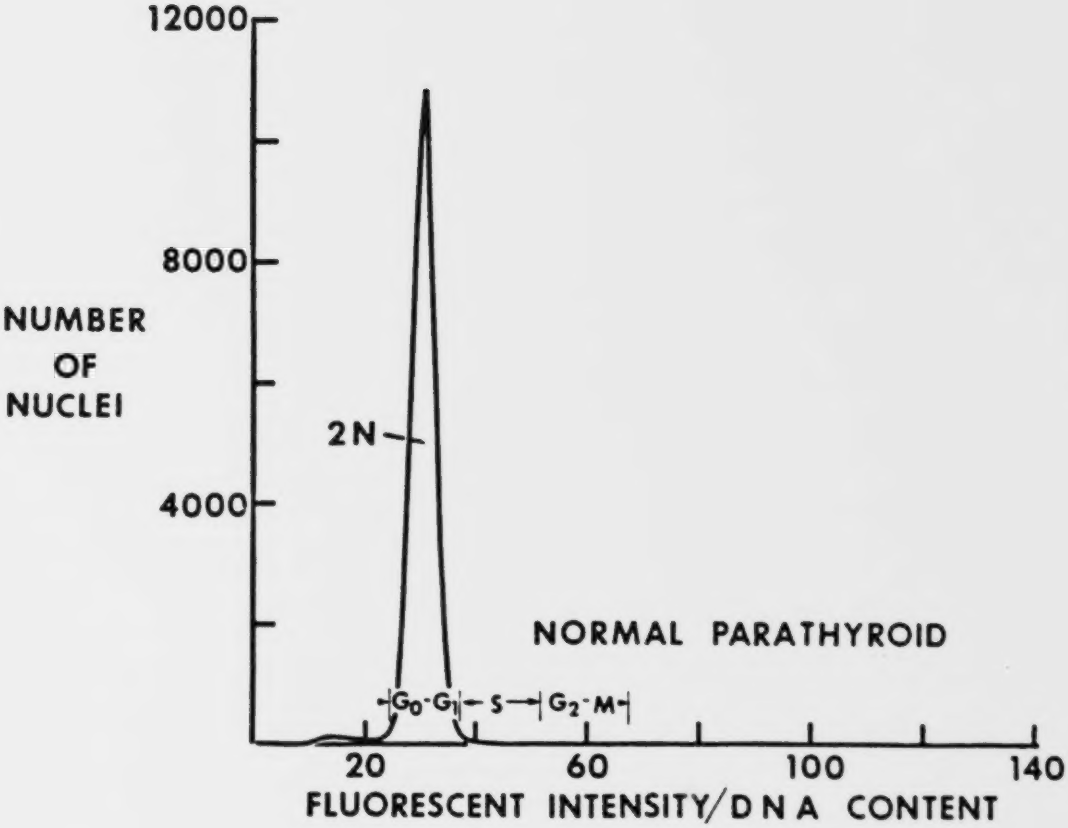

12000
8000
4000
NUMBER OF NUCLEI
2N
G$_0$-G$_1$
S
G$_2$-M
NORMAL PARATHYROID
20
60
100
140
FLUORESCENT INTENSITY/D N A CONTENT

Figure 9.3-2: Typical DNA histogram from an adenomatous parathyroid. The three peaks represent diploid, tetraploid and octaploid nuclei.

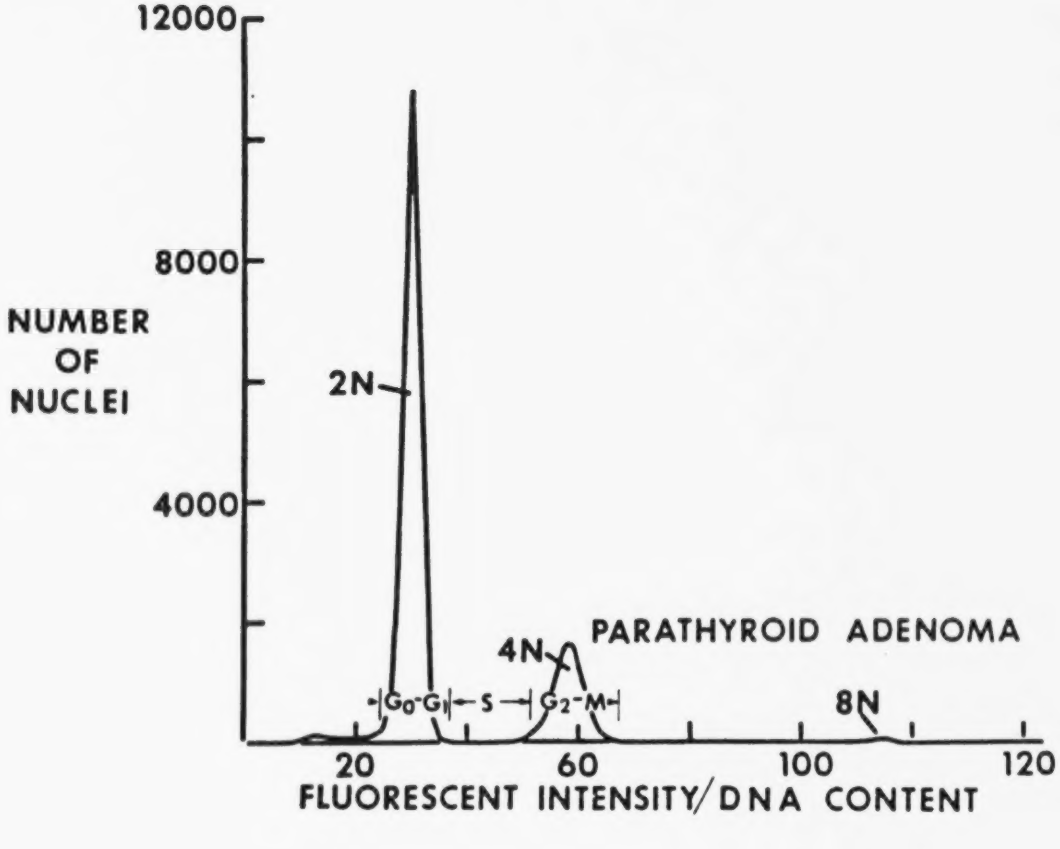

12000
8000
4000
NUMBER
OF
NUCLEI
2N
4N
8N
PARATHYROID ADENOMA
G0-G1
S
G2-M
20
60
100
120
FLUORESCENT INTENSITY/DNA CONTENT

Figure 9.3-3: Comparison of averaged normal and adenomatous parathyroid histograms. In the tetraploid region (channels 50-70) a significant difference between these two groups was observed ($p<.04$).

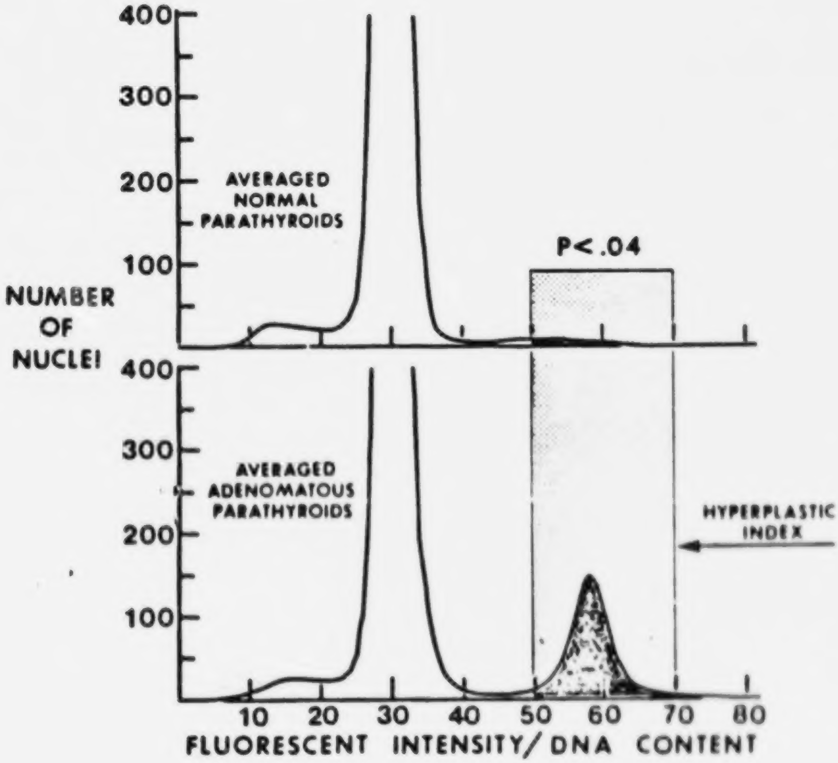

NUMBER
OF
NUCLEI
400
300
200
100
AVERAGED
NORMAL
PARATHYROIDS
P<.04
AVERAGED
ADENOMATOUS
PARATHYROIDS
HYPERPLASTIC
INDEX
10
20
30
40
50
60
70
80
FLUORESCENT INTENSITY/DNA CONTENT

Figure 9.3-4: Hyperplastic index for each normal, "normal", hyperplastic and adenomatous parathyroid sample. The hyperplastic index is in log scale because the samples are normally distributed about a mean value when in the log domain. The height of the bars represent mean values and the range bars represent plus/minus one standard deviation. The hyperplastic and adenomatous parathyroid groups were significantly different from normal ($p<.02$ and $p<.04$ respectively).

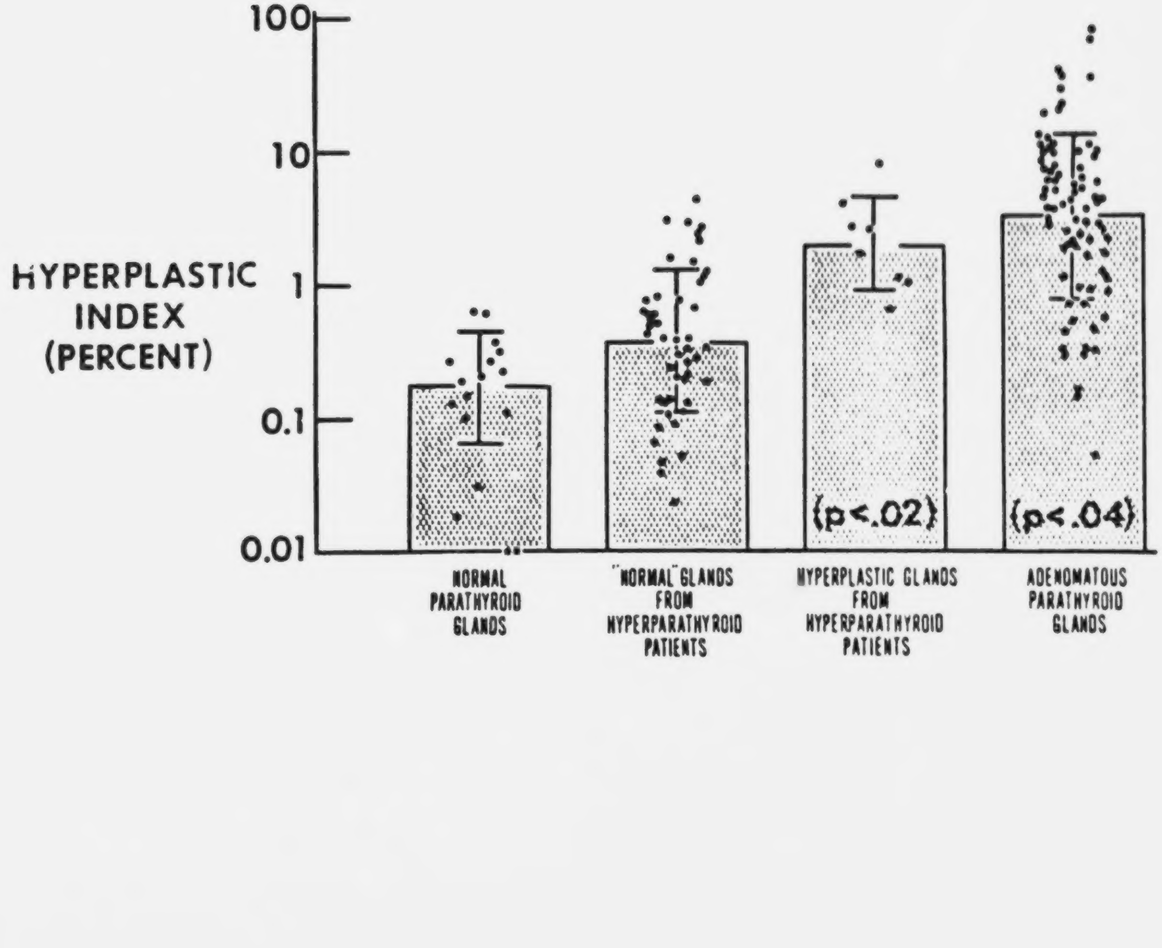

100
10
1
0.1
0.01
HYPERPLASTIC INDEX (PERCENT)
(p<.02)
(p<.04)
NORMAL PARATHYROID GLANDS
"NORMAL" GLANDS FROM HYPERPARATHYROID PATIENTS
HYPERPLASTIC GLANDS FROM HYPERPARATHYROID PATIENTS
ADENOMATOUS PARATHYROID GLANDS

Figure 9.3-5: Probability density functions for normal and adenomatous hyperplastic indices. The decision axis at the point of intersection of these two curves is represented as a dashed line.

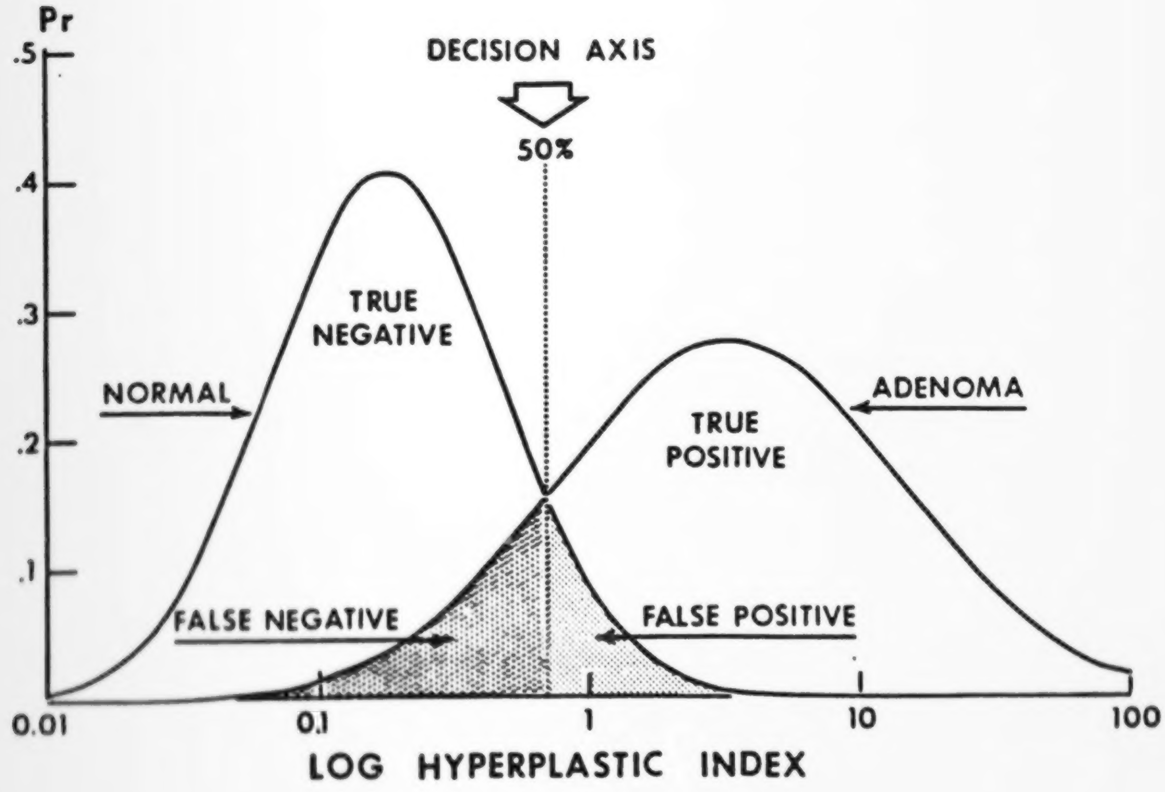

Pr
.5
.4
.3
.2
.1
DECISION AXIS
50%
TRUE NEGATIVE
NORMAL
ADENOMA
TRUE POSITIVE
FALSE NEGATIVE
FALSE POSITIVE
0.01
0.1
1
10
100
LOG HYPERPLASTIC INDEX

Figure 9.3-6: The calculated percent chance of being an adenoma for each normal, "normal", hyperplastic and adenomatous parathyroid sample. Heights of the bars represent mean values for each group. Samples falling below the 50% decision axis are negative and above are positive for being in the adenomatous range.

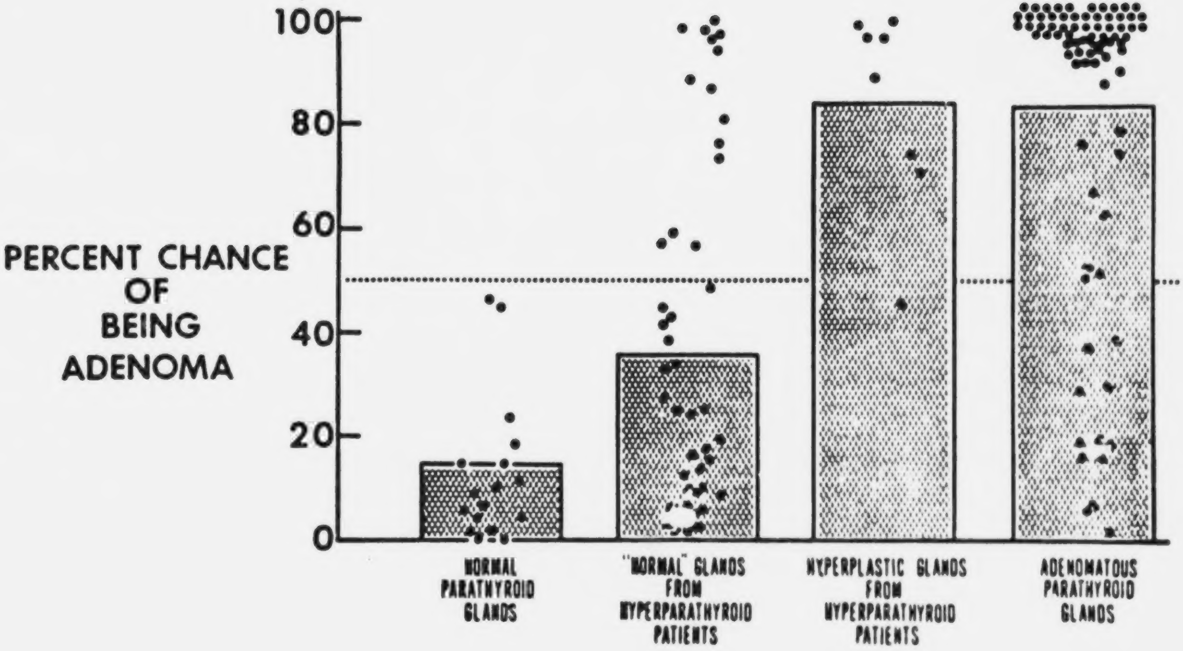

100
80
60
40
20
0
PERCENT CHANCE OF BEING ADENOMA
NORMAL PARATHYROID GLANDS
"NORMAL" GLANDS FROM HYPERPARATHYROID PATIENTS
HYPERPLASTIC GLANDS FROM HYPERPARATHYROID PATIENTS
ADENOMATOUS PARATHYROID GLANDS

Figure 9.3-7: Parathyroid classification summary. The computer predicted classification is compared to each histological tissue diagnosis.

PARATHYROID CLASSIFICATION SUMMARY

PERCENT CHANCE OF BEING ADENOMA		NORMAL PARATHYROID GLANDS	"NORMAL" GLANDS FROM HYPERPARATHYROID PATIENTS	HYPERPLASTIC GLANDS FROM HYPERPARATHYROID PATIENTS	ADENOMATOUS PARATHYROID GLANDS
	≥50	0 (0%)	14 (29%)	7 (87%)	68 (85%)
	<50	17 (100%)	34 (71%)	1 (13%)	12 (15%)

TISSUE DIAGNOSIS

10. References

1. Horan, P: Optical Techniques in Cell Enumeration. McGraw-Hill Yearbook of Science and Technology, 1976.
2. Brunsting, A et al: Coulter Two Parameter Cell Sorter System. ed Mendelson M, Melamed M, and Mullaney P. Wiley, New York, 1979.
3. Latimer P, Brunsting A, Pyle BE, and Moore C: Effects of Asphericity on Single Particle Scattering. Applied Optics. 17:3152, 1978.
4. Miller F: COLLEGE PHYSICS. Harcount, Brace and World, Inc, New York. Pg 577, 1967.
5. Driscoll WG and Vaughon WV: NONIMAGING DETECTORS. McGraw Hill, New York. 4:7, 1978.
6. Gill J: Methods course in fluorescent flow cytometry-measurements for DNA analysis. University of Rochester, Department of Pathology, Rochester, New York, October 27, 1978.
7. Horan P, Kappler JW: Automated fluorescent analysis for cytotoxicity assay. J of Immun Meth 18:309-316, 1977.
8. Crissman HA and Steinkamp JA: Rapid, simultaneous measurement of DNA, protein, and cell volume in single cells from large mammalian cell populations. J Cell Biol 59:766, 1973.
9. Krishan A: Flow cytometric long term storage of

propidium iodide/ citrate-stained meterial. Stain Tech 52:339,343, 1977,

10. Krishan, A :Rapid Flow-Cytofluorometric Analysis of Mammalian Cell Cycle by Propidium Iodide Staining. J Cell Biol 66:763, 1975.
11. Vindelov LL: Flow microfluorometric analysis of nuclear DNA in cells from solid tumors and cell suspensions. Virchows Arch B Ell Path 24:227, 1977.
12. Langman J: MEDICAL EMBRYOLOGY. The Williams and Wilkins Company, Baltimore. Chapt 1:46, 1972.
13. Dean PN and Anderson EC: The rate of DNA synthesis during S phase by mamalian cells in vitro. Proceedings of the International Symposium, The Netherlands, 1974.
14. Hartman NR and Cristensen IJ: A theory connecting the DNA synthesis rate, cell flux, cell loss, and the deconvoluted DNA distribution measured with flow cytometry. Third annual meeting for the Society of Cell Kinetics. Held in Miami, March, 1979.
15. Read ME: Investigation of the Working Principles of Tomographic Devices and their Imaging Quality. University of Miami Ph.D. Thesis. Pgs 40-50, 1973.
16. Ash RB: INFORMATION THEORY. Interscience Publishers, New York. Pgs 21-24, 1966.
17. Meyer SL: DATA ANALYSIS FOR SCIENTISTS AND ENGINEERS. John Wiley and Sons, New York. Pg 115, 1975.

18. Kaiser JF and Reed WA: Data Smoothing using low-pass digital filters. Rev Sci Instrum 48:1447, 1977.

19. General Electric Numerical Analysis Routines Manual 5700.02C, 1974.

20. Leary JF, Todd P, Wood JC, and Jett J: Laser flow cytometric lightscatter and fluorescent pulse width and pulse rise time sizing of mammalian cells J of Hist Cyto. 27:315-320, 1979.

21. Drapes NR and Smith H: APPLIED REGRESSION ANALYSIS. John Wiley and Sons, Inc. New York. Pgs 263-304, 1966.

22. Marquardt, DW: An algorithm for least-squares estimation of nonlinear equations. J Soc and Appl Math 11:431-441, 1963.

23. Pearlman AL: Information content of flow microfluorometric histograms: analysis and applicatins. 1976 Quarterly Laboratory of Chemical Biodynamics, Lawrence Livermore Laboratory, 1976.

26. Brunsting A: Unpublished results.

27. Brunsting A, Collins J, Kain ,Bagwell, CB: An Examination of Some Basic Assumptions in DNA Histogram Analysis. J of Cell Kinet, 1978.

28. Dean PN and Jett JH: Mathematical Analysis of DNA distributions Derived from Flow Microfluorometry. J Cell Biol 60:523, 1974.

29. Fried, J: Mathematical Analysis of Flow Microfluorometric Data and its Application to the Determination of Cell Cycle Parameters. Mathematical Models in Cell Kinetics, edited by Valleron, AJ European Press Medikon, 1975.

30. Fried, J: Analysis of Deoxyribonucleic Acid Histograms for Flow Cytometry, Estimation of the Distribution of Cells in S Phase. J Hist Cyto. 25:942. 1975.

31. Jett JH: Mathematical Analysis of DNA Histograms from Asynchronous and Synchronous Cell Populations. Proceedings of the Third International Symposium on Pulse-Cytophotometry, held in Vienna, Austria (March 30-April 1, 1977); European Press, Ghent, Belgium.

32. Valkenburg ME: NETWORK ANALYSIS. Prentice-Hall, inc. Chapt 8:222-229, 1955

33. Abramowitz M and Stegun I: HANDBOOK OF MATHEMATICL FUNCTIONS. National Bureau of Standards, Washington DC. Chapt 7:299, 1964.

34. Bagwell CB, Hudson JL, and Irvin GL: Nonparametric Flow Cytometry Analysis. J of Hist Cyto 25:293, 1979.

35. Lusted LB: Introduction to Medical Decision Making. Charles Thomas Publisher, Illinois, 1968

36. Meyer SL: Data Analysis. John Wiley and Sons, Inc., New York, 1975.

37. Pollack A, Bagwell CB, Hudson JL, Irvin GL, III:

Differences in flow cytometry and ^{3}H-TdR analyses of perturbed human lymphoctyes. J Histochem Cytochem 26:000,1978 (in press)

38. Young IT: Proof without prejudice. J Histochem Cytochem 25:935, 1977.

39. Schwartz, S et al. PRINCIPLES OF SURGERY: Thyroid and parathyroid. McGraw Hill Book Comp. Pg 1545, 1979.

40. Robins SL: PATHOLOGIC BASIS OF DISEASE. WB Saunders Co, Philadelphia. Pg 1349, 1974.

41. Irvin, GL and Bagwell, CB: Flow cytometry detection of parathyroid hyperplasia. American Journal of Surg, 1979 (submitted).

42. Edis AJ, Beahrs OH, van Heerden JA, Akwari OE: "Conservative" versus "liberal" approach to parathyroid neck exploration. Surgery 82:466, 1977.

43. Clark OH, Way LW, Hunt TK: Recurrent hyperparathyroidism. Ann Surg 184:391, 1976.

44. Coffey RJ, Lee TC, Canary JJ: The surgical treatment of primary hyperparathyroidism: a 20 year experience. Ann Sur 185:518, 1977.

45. Haff RC, Ballinger WF: Causes of recurrent hypercalcemia after parathyroidectomy for primary hyperparathyroidism. Ann Surg 173:884, 1971.

46. Paloyan E, Lawrence AM: The rationale for subtotal parathyroidectomy. Controversy in Surgery (Varco RL, Delaney JP, eds). Philadelphia, WB Saunders, 1976.

47. Attie JN, Wise L, Mir R, Ackerman LV: The rationale against routine subtotal parathyroidectomy for primary hyperparathyroidism. Am J Surg 136:437, 1978.

48. Wang C, Rieder SV: A density test for the intraoperative differentiation of parathyroid hyperplasia from neoplasia. Ann Surg 187:63, 1978.

49. Bengtsson A, Grimelius L, Johansson H, Ponten J: Nuclear DNA-content of parathyroid cells in adenomas, hyperplastic and normal glands. ACTA Microbiol Scand 85:455, 1977.

50. Fialkow PJ, Jackson CE, Block MA, Greenawald KA: Multicellular origin of parathyroid "adenomas". NEJM 297:696, 1977.

51. Collins JM, Berry DE, and Bagwell CB: Different rates of DNA synthesis during the S-phase of log-phase Hela S3, WI-38, and 2RA cells. J of Mol Bio, 1979. (in prep).

52. Bevington P :Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill Book Co., New York, 1969.

VITA

Charles Bruce Bagwell was born on June 29,1948 in McKeesport, Pennsylvania. His parents are Kathryn Hunsberger Bagwell and Charles Ernest Bagwell. He graduated from Staples High School in 1967. He majored in Chemistry at the University of Connecticut and received his B.A. in 1972. After a year of virology research in Phillip Marcus' lab in Connecticut and a year of immunology research in George Irvin's laboratory in Miami, he was accepted into the University of Miami Medical School in 1973. He entered the M.D. to Ph.D. program in 1976 and received his M.D. degree in 1977. In August of 1979 he was granted the degree of Doctor of Philosophy by the University of Miami, Coral Gables, Florida.

Permanent address: 1640 N.W. 19th Avenue
Miami, Florida 33125